修士論文

原言語と目標言語の反復的データ拡張に基づく
低資源言語のニューラル機械翻訳

花田 佳文

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和７年３月

Abstract

In recent neural machine translation, which is the mainstream of current machine translation research, high translation accuracy is achieved by training translation models using large parallel corpora. However, in low-resource languages where available language resources are limited, the application of state-of-the-art neural machine translation technology has been behind. Nevertheless, several attempts have been made to improve neural machine translation performance for low-resource languages. For example, Iterative Back-Translation is a method that iteratively enhances translation performance by repeating training of a back-translation model (a model of translation from a target language to a source language) and expansion of a parallel corpus using this back-translation model. However, it has not been sufficiently explored whether such existing methods are effective for translation in all low-resource languages.

This research focuses on machine translation from Hokkaido Ainu, one of the extreme low-resource languages in Japan, to Japanese, and aims to enhance its performance. The existing Iterative Back-Translation is first applied to Ainu-Japanese translation and its effectiveness is empirically verified. Furthermore, we propose "Iterative Cyclic-Back-Forward-Translation" (hereafter we call it "Iterative Cyclic-Translation" for simplicity) as an improvement to Iterative Back-Translation.

First, we built an Ainu-Japanese parallel corpus by extracting texts from several documents. The Ainu language materials used for the construction of this parallel corpus were limited to Hokkaido Ainu. As of 2024, there is no standardized orthography for Hokkaido Ainu, and spelling and morpheme segmentation are different across documents, so the notation of Ainu sentences in the documents is manually converted into the uniform notation when they are compiled in the parallel corpus. Finally, we constructed a parallel corpus consisting of 23,337 translation pairs, which serves as our initial parallel corpus.

Our proposed Iterative Cyclic-Translation is carried out as follows. First, data augmentation is performed by cyclic translation on the initial parallel corpus. Sentences in the target language (Japanese) in the parallel corpus are translated to sentences in an auxiliary language and then translated back to Japanese, resulting in generating paraphrases of the sentences in the target language. These paraphrases are combined with the original sentences in the source language (Ainu) to acquire new translation pairs, which are then added to the parallel corpus. In this research, English is chosen as the auxiliary language, as Japanese-English machine translation is well-developed and existing high-performance translation systems are available. Next, data augmentation is performed by back-translation. A back-translation model is trained using the augmented parallel corpus and is used to translate target language sentences into source language sentences. New trans-

lation pairs obtained by this procedure are added to the parallel corpus. Then, data augmentation is performed by forward translation. The parallel corpus is further augmented by training a translation model using the augmented parallel corpus and translating source language sentences into target language sentences using this translation model. The above procedures are repeated several times to obtain the final Ainu-to-Japanese translation model. In the existing Iterative Back-Translation, data augmentation by the back-translation, which generates new source language sentences, and data augmentation by the forward translation, which generates new target language sentences, are repeated. In contrast, in the proposed Iterative Cyclic-Translation, additional data augmentation by the cyclic translation, which generates new target language sentences, is incorporated. Throughout the iterative training process, the addition of cyclic-translation-based data augmentation results in a larger final parallel corpus compared to Iterative Back-Translation, which is expected to improve the translation performance of the machine translation model trained on it.

In the evaluation experiments, three models were compared: an Ainu-to-Japanese translation model trained with the initial parallel corpus only (baseline model), a model trained with Iterative Back-Translation, and a model trained with Iterative Cyclic-Translation. For both Iterative Back-Translation and Iterative Cyclic-Translation, the maximum number of iterations was set to 2. Ainu test sentences were translated into Japanese and the accuracy of the translation was evaluated using BLEU and CHRF^{++} metrics.

As for the model trained with Iterative Back-Translation, after one iteration, the BLEU and CHRF^{++} were improved by 8.95 and 6.48 points, respectively, compared to the baseline model. However, after two iterations, the improvements were only 0.66 points for BLEU and 1.47 points for CHRF^{++}, which were lower than the scores achieved by the model after one iteration. On the other hand, when applying our proposed Iterative Cyclic-Translation method, there was no improvement in either BLEU or CHRF^{++} scores compared to the baseline after one and two iterations.

Furthermore, we conducted manual evaluation of several sentences. The sentences translated by the Iterative Back-Translation model after one iteration were generally comprehensible and acceptable. However, the model using Iterative Cyclic-Translation demonstrated a tendency to produce mistranslations of Ainu cultural terms and omissions of low-frequency words. Regarding the translation of the Ainu folklore specific expressions such as parallelism and periphrasis, while all models could generate translations that captured the essential meaning, the model obtained by Iterative Cyclic-Translation often generated simplified expressions by inadequately summarizing redundant expressions in the folklore style.

In automatic evaluation, our proposed Iterative Cyclic-Translation showed lower

BLEU and CHRF^{++} scores compared to the existing Iterative Back-Translation. This may be because the existing Japanese-English machine translation system used in cyclic translation is inappropriate for translating sentences containing Ainu cultural terms, resulting in generating low-quality augmented translation pairs. Additionally, the current initial Ainu-Japanese parallel corpus is both qualitatively and quantitatively insufficient. The number of Ainu language documents used in this research was limited, and many Japanese sentences in the initial parallel corpus were not easy to understand. Furthermore, to prepare translation pairs as much as possible, we created a unified parallel corpus that consists of Ainu sentences in different domains, such as folktales and colloquial styles, and different dialects. Training a single translation model that handles different domains and dialects may have resulted in the failure to train a sophisticated translation model.

In future work, in addition to expanding and improving the documents included in the Ainu-Japanese parallel corpus, we aim to add English translations for Ainu-Japanese translation pairs, train models between target and auxiliary languages for improvement of cyclic translation, and verify whether translation accuracy is improved. Furthermore, given that folktales and colloquial texts exhibit differences in expressions and vocabulary across domains and dialects, we verify whether our proposed method achieves similar improvements in translation accuracy when a sufficient parallel corpus is prepared for each dialect using more Ainu documents. Additionally, to ensure the quality of augmented Ainu sentences, we plan to explore a method that incorporates a grammatical check of Ainu sentences by verifying whether the number of arguments of a verb in a sentence is correct.

概要

近年の機械翻訳研究の主流であるニューラル機械翻訳では，大量のパラレルコーパスを用いて翻訳モデルを学習し，高い翻訳精度を実現している．一方，低言語資源の言語では利用可能な言語が限られており，最先端のニューラル機械翻訳技術の適用が遅れている．とはいえ，低資源言語のニューラル機械翻訳の性能を高める研究も行われている．例えば，反復的逆翻訳は，目標言語から原言語への翻訳モデル (逆翻訳モデル) の学習とそれによる翻訳を利用したパラレルコーパスの拡張を繰り返すことで漸進的に翻訳の性能を高める手法である．しかし，このような先行研究の手法がどのような低資源言語の翻訳についても有効であるかは十分に探究されていない．

本研究では，低資源言語の一つである北海道アイヌ語から日本語への機械翻訳を対象とし，その性能を高めることを目的とする．先行研究の反復的逆翻訳をアイヌ語から日本語への翻訳に適用し，その有効性を検証する．さらに，反復的逆翻訳を改良した「反復的巡回・逆・順翻訳」(以下，単に「反復的巡回翻訳」) を提案する．

まず，いくつかの文献からアイヌ語・日本語のテキストを抽出し，アイヌ語・日本語パラレルコーパスを構築する．このパラレルコーパスで用いるアイヌ語文献は北海道アイヌ語のものに限った．アイヌ語は 2024 年現在正書法が定義されておらず，文献によって綴りや形態素区切りの揺れがあるため，パラレルコーパス整備時に人手で表記を統一した．最終的に 23,337 の翻訳ペアからなるパラレルコーパスを構築した．これを初期のパラレルコーパスとする．

本研究で提案する反復的巡回翻訳の処理の流れは以下の通りである．まず，初期のパラレルコーパスに対し，巡回翻訳によるデータ拡張を行う．パラレルコーパスにおける目標言語の文 (日本語文) を補助言語に翻訳し，それを日本語に再翻訳することで別の文に言い換える．これを元の原言語の文と組み合わせることで新たな翻訳ペアを獲得し，パラレルコーパスに追加する．本研究では補助言語として英語を用いる．高資源言語である日英機械翻訳は研究が進んでおり，既存の高性能の翻訳システムが利用できるためである．次に，逆翻訳によるデータ拡張を行う．拡張したパラレルコーパスを用いて逆翻訳モデルを学習し，これを用いて目標言語の文を原言語の文に翻訳することで，新たな翻訳ペアを得て，パラレルコーパスに追加する．続いて，順翻訳によるデータ拡張を行う．拡張したパラレルコーパスを用いて翻訳モデルを学習し，これを用いて原言語の文を目標言語の文に翻訳することでパラレルコーパスを拡張する．この一連の操作を反復することで最終的なアイヌ語から日本語への翻訳モデルを得る．先行研究の反復的逆翻訳では原言語への逆翻訳によるコーパス拡張と目標言語への順翻訳によるコーパス拡張を繰り返すが，反復的巡回翻訳ではこれに加えて目標言語の巡回翻訳によるコーパス拡張の処理が加わる．反復学習全体で，巡回翻訳によるコーパス拡張が追加されることで，反復的逆翻訳よりも最終的に得られるパラレルコーパスの量が多くなり，それから学習する翻訳モデルの翻訳精度の向上が期待できる．

評価実験では，初期パラレルコーパスのみで学習したモデル，反復的逆翻訳によって学習したモデル，反復的巡回翻訳によって学習したモデルを比較した．反復的逆翻訳と反復的巡回翻訳では最大の反復回数を 2 回とした．アイヌ語テスト文を日本語文へ翻訳し，翻訳精度を BLEU と chrF^{++} を指標として評価した．反復的逆翻訳によって，反復回数が 1 回の時はベースラインモデルと比べて，BLEU が 8.95 ポイント，chrF^{++} が 6.48 ポイント向上した．しかし，反復回数が 2 回の時は BLEU が 0.66 ポイント，chrF^{++} が 1.47 ポイントの向上となり，反復回数が 1 回の時のモデルよりもスコアは低かった．一方，本研究の提案手法である反復的巡回翻訳を適用したとき，反復回数を増やすたびに BLEU スコア，chrF^{++} スコアともに低下し，ベースラインモデルや反復的逆翻訳を適用したモデルと比べて改善はみられなかった．さらに，いくつかの文に対して人手評価を行った．反復回数が 1 回の時の反復的逆翻訳モデルによって翻訳された文は概ね理解可能な訳出であり許容範囲であった．一方で，反復的巡回翻訳を適用したモデルはアイヌ文化関連語の誤訳や低頻度語の訳出漏れが確認された．対句表現や迂言的な表現といった雅語特有の表現の翻訳については，いずれのモデルも文義は捉えていたものの，反復的巡回翻訳を適用した際のモデルは雅語の冗長な表現がまとめられて淡白な表現の文が出力されたことが確認された．

反復的巡回翻訳によって精度が向上しなかった原因として，巡回翻訳で用いた既存の日英機械翻訳システムがアイヌ文化関連語を含む文の翻訳に適しておらず，拡張した翻訳ペアの品質が低かったことが考えられる．また，アイヌ語・日本語パラレルコーパスの質的・量的な問題もある．本研究で利用したアイヌ語の文献数が少なく，また初期パラレルコーパスの日本語文もわかりにくいものが多かった．さらに，データ数を確保するために口承文芸や会話体などのドメインや方言を区別せずに一つのパラレルコーパスを作成したが，異なるドメイン・方言を単一の翻訳モデルとして学習したことで，結果として適切な翻訳モデルが学習できなかった可能性がある．

今後，アイヌ語・日本語パラレルコーパスに含まれる文献の拡張・改良に加え，補助言語資料として英語訳を拡張し，目標言語・補助言語間の翻訳モデルを学習したとき，原言語から目標言語方向への翻訳精度が向上するかを検証を行いたい．また，口承文芸と会話文のドメイン，方言によって表現や語に差異があるため，文献数を確保することにより，方言ごとに分けた場合でもある程度のデータ数が確保できたとき，本提案手法によって翻訳精度が同様に向上するか検証を行いたい．これに加え，翻訳後のアイヌ語文の品質を担保する一案としてアイヌ語文の非文検査を行い，文中の動詞から取りうる項数を割り出し，項数を検査する手法を実施したい．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年の機械翻訳の研究はニューラル機械翻訳が主流であるが，一般に大量のパラレルコーパスを必要とするため，利用可能な言語データが少ない低資源言語については最先端のニューラル機械翻訳技術の適用が遅れている．低言語資源を対象にニューラル機械翻訳の性能を向上させることを試みた研究もあるが，その翻訳精度は十分に高くなく，改善の余地がある．

機械翻訳の研究がそれほど進んでいない低言語資源の言語のひとつにアイヌ語がある．アイヌ語は日本の先住民族言語の一つであり，北海道，樺太，千島で使用され，かつては本州北部でも話されていた言語である．SOV 型の言語で統語的に日本語と類似する点もある一方で，抱合的特徴を有し接頭辞を多用するなどの異なる点も多く，基本的には日本語とは異なる独立した言語である [3]．アイヌ語の系統は明確に定まったものがなく，孤立言語とされている．北海道アイヌ語，樺太アイヌ語，千島アイヌ語を一つの語族とみなし，アイヌ語族 (Ainuic) とする場合もある [32]．樺太アイヌ語，千島アイヌ語は 20 世紀に消滅したとされ，北海道アイヌ語も母語話者はいなくなっているとされる [36]．2024 年現在，アイヌ語はいわゆる共通語や標準語と呼ばれる言語変種が存在せず，また方言間での差異が存在する．図 1.1 はアイヌ語方言の分布を示している．この図に示すように，北海道アイヌ語は北東部方言と南西部方言に大別され [3]，南西部方言の沙流方言，幌別方言，千歳方言や北東部方言の静内方言，旭川方言などの地域ごとに異なる特徴を持っており，単語や人称などに地域差が見られる．このうち，南西部方言はアイヌ語の中で最も文献が記録されている方言である．

アイヌ語の各方言はアイヌ民族にルーツのある人達を中心に復興・継承活動が続けられており，アイヌ語弁論大会や北海道各地並びに東京でアイヌ民族に向けたアイヌ語教室が実施されている．その一方で，一般の学習者が利用できる機械翻訳などの技術は学術的な研究を除くと筆者の知る限り存在しない．このような現状を踏まえ，アイヌ語の復興・継承活動，利用や学習機会の増加，またアイヌ語話者の言語権行使のために，手軽に利用できるアイヌ語と日本語の機械翻訳システムが求められる．

なお，これ以降本論文では特筆がない限り，北海道アイヌ語を「アイヌ語」と呼称する．

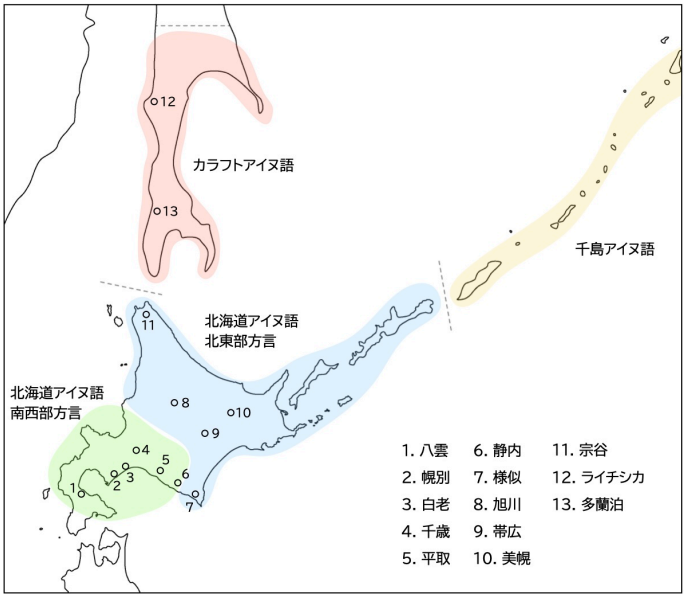

図 1.1: アイヌ語の方言分布

1.2 目的

本研究では，原言語であるアイヌ語から目標言語である日本語への翻訳を対象とした新しいニューラル機械翻訳の方法を提案する．原言語文と目標言語文の拡張と翻訳モデルの学習を繰り返す反復的逆翻訳 (Iterative Back Translation) の手法をアイヌ語→日本語 NMT (Neural Machine Translation) モデルに適用し，機械翻訳の精度向上を目指す．これに加え，目標言語文を原言語以外の補助言語を用いた巡回翻訳を経て別の文に言い換え，新しい翻訳ペアを獲得することでパラレルコーパスを拡張する手法を提案する．なお，本論文では巡回翻訳で使用する補助言語は英語とする．

逆翻訳 (Back-translation) とは，ある言語の文を別の言語の文に翻訳し，再度それを元の言語の文に翻訳することで，パラレルコーパスを拡張する手法である [27]．逆翻訳は低言語資源環境下でのコーパス拡張 (Corpus Augmentation) の手法として使用されており [8, 9]，同様に低言語資源であるアイヌ語・日本語パラレルコーパスの拡張にも有用であると考えられる．巡回翻訳も逆翻訳と同様に別の言語への翻訳と元の言語への再翻訳を経てコーパス拡張を行う手法であるが，翻訳対象とする原言語，目標言語以外の言語を補助言語として用いる点が異なる．

本研究で取り扱う低言語資源の言語はアイヌ語とするが，同様に話者の少ない小言語 (その多くは低言語資源言語) を対象にした精度の高い機械翻訳モデルの訓練方法を整備することができれば，その言語の利用や学習の促進の一助になるであろう．また当該言語話者・学習者を対象に，個人の自由な言語利用を制限せずに，大言語話者が利用できるデジタル環境や最先端技術と同様の環境が整備されるべきであり，本研究の成果は小言語話者・学習者の言語権向上に貢献すると考

えている．

1.3　論文の構成

本論文の構成は以下のとおりである．2 章では，本論文に関連する研究を概観する．3 章では本研究で利用するアイヌ語・日本語パラレルコーパスについて説明する．4 章ではアイヌ語から日本語への翻訳モデルを学習する提案手法の詳細について述べる．5 章では提案手法の評価実験について報告する．最後に，6 章では本稿のまとめと今後の課題について述べる．

第2章　関連研究

本章では逆翻訳，コーパス拡張やアイヌ語を対象とした自然言語処理に関する先行研究のうち，本提案手法に特に関連の深い研究について述べる．

2.1　逆翻訳

逆翻訳　Sennrich らは，ニューラルネットワークのアーキテクチャを変更せず，NMT モデルの翻訳精度を向上させる手法として，目標言語の単言語データを原言語へ逆翻訳 (Back-translation) し，獲得した合成データを原言語から目標言語方向のニューラル機械翻訳モデルの学習データに追加する手法を提案した [27]．評価実験では，英語・ドイツ語間の双方向翻訳タスク，ならびにトルコ語から英語の翻訳タスクにこの手法を適用し，既存手法に比べ全ての翻訳タスクで最新性能 (State-of-the-Art) を達成し，データ拡張によって過学習の抑制と流暢さの向上が確認できたと考察している．

反復的逆翻訳　Hoang らは Sennrich らの逆翻訳を発展させ，これを反復的に実施する反復的逆翻訳 (Iterative Back-translation) を提案した [9]．

この手法では，まず対訳コーパス Dp を使用して，目標言語から原言語方向の NMT モデルを訓練する．そのモデルを使って目標言語の単言語コーパスの文を翻訳し，得られた擬似原言語文 $\hat{s}$ と目標言語文 t を新たな対訳ペア S として初期対訳コーパス Dp に追加する．次に，この拡張後の対訳コーパスを使用して，原言語から目標言語方向の NMT モデルを訓練する．そのモデルで原言語の単言語コーパスの文を翻訳し，その出力として新たに目標言語文 $\hat{t}$ を得る．擬似目標言語文 $\hat{t}$ を原言語文 s と組み合わせて，新たな対訳ペア S' として初期対訳コーパス Dp に再度追加する．この一連の処理を収束条件を満たすまで実施し，最終的に反復的に再学習された NMT モデルを獲得する．この反復的逆翻訳の擬似アルゴリズムを Algorithm 1 に示す．

Algorithm 1 反復的逆翻訳 [9]

Input: original parallel corpus D^p, monolingual source D^s, and target D^t text

Let $T_{\leftarrow} = D^p$
repeat
  Train target-to-source model $\Theta_{\leftarrow}$ on $T_{\leftarrow}$
  Use $\Theta_{\leftarrow}$ to create $S = \{\hat{s}, t\}$, for $t \in D^t$
  Let $T_{\rightarrow} = D^p \cup S$
  Train source-to-target model $\Theta_{\rightarrow}$ on $T_{\rightarrow}$
  Use $\Theta_{\rightarrow}$ to create $S' = \{s, \hat{t}\}$, for $s \in D^s$
  Let $T_{\leftarrow} = D^p \cup S'$
until convergence condition reached

Output: newly-updated models $\Theta_{\leftarrow}$ and $\Theta_{\rightarrow}$

この手法は，言語資源量が潤沢な環境でも，初期のパラレルコーパスの単語数が 100,000 件程度の低言語資源の環境でも翻訳性能が向上したと報告されている．また，Sennrich らの研究でも言及されていたが [27]，初期コーパスに追加する新しい対訳ペアの量の調整も必要であると述べている．Hoang らの研究では，低言語資源環境として利用した英語・ペルシア語の翻訳タスクにおいて，初期コーパスと拡張コーパスを 1:2，もしくは 1:3 の割合で訓練に使用した時，1:1 の割合でコーパスを拡張した時と比べて翻訳精度が向上したと報告している．

2.2 アイヌ語を対象とした自然言語処理研究

ルールベース機械翻訳，統計的対訳語抽出 桃内らは，アイヌ語から日本語の機械翻訳の精度向上を目的に，構文上の類似性の近さから直接単語をアイヌ語から日本語に翻訳するのではなく，より自然な日本語文に変換するための変換処理について考察を行った [19]．ルールベース機械翻訳モデルの検討であるが，アイヌ語・日本語翻訳の研究の萌芽である．

越前谷らは，アイヌ語―日本語対訳コーパスからの対訳語の自動抽出の手法として，局所着目方式 (LFM: Local Focus Method) を提案した [6]．この手法では対訳文中の共通部分に注目し抽出する名詞対訳語を決定する．2 つの共通部分間の差分から名詞を特定する方法と，1 つの共通部分の周辺の差分から名詞を特定する方法の 2 通りが提案された．281 種類の名詞対訳語を対象とした評価実験の結果，再現率 60.1%，適合率 70.1%で名詞対訳対が抽出されたと報告している．

また越前谷らは，小規模な対訳コーパスのみでも語の対応関係の自動抽出を可能にする方法として局所着目型学習 (LFL: Local Focus-based Learning) を提案した [7]．LFL は，対訳文中の共通部分とその隣接部分を局所部分として，対応語を探索抽出する「対訳文ペア方式」と，局所部分から対訳語を自動抽出するために

獲得したコロケーションテンプレートと呼ばれるルールを対訳ペアに適用し，対訳語を抽出する「テンプレート方式」を組み合わせたルールベースの対訳語自動抽出手法である．なお，この手法は局所着目型「学習」と呼ばれているが，出現頻度などに依存する統計的な手法ではない．288 文からなるアイヌ語―日本語対訳コーパスに対して局所着目型学習を適用したところ，再現率 54.0%，適合率 60.8% の精度で名詞対訳語と動詞対訳語が抽出できた．対訳コーパスに対して統計的に処理を行い対訳語の組を抽出する既存手法と比べ，再現率が 10%高かったと報告している．

ニューラル機械翻訳　Miyagawa は，ニューラル機械翻訳モデルである MarianMT を用い，インターネット上に公開されているアイヌ語―日本語コーパスをデータセットとして，日本語→アイヌ語翻訳モデル，アイヌ語→日本語翻訳モデル，ア・日双方向翻訳モデルをそれぞれ訓練したところ，BLEU スコアで日本語→アイヌ語で 32.90 と高い精度を達成し，またアイヌ語→日本語で 10.45 を，双方向翻訳モデルで 29.91 をそれぞれ達成したと報告している [18]．なお，ア・日双方向翻訳モデルは日本語→アイヌ語，アイヌ語→日本語の翻訳の両方に適用できる．すなわち両方向の翻訳を同一のモデルで実現している．アイヌ語→日本語方向の翻訳については，アイヌ語の言語資源が限られていたために高い BLEU スコアが得られなかったのではないかと考察している．

また，Igarashi らは，系列変換モデルである T5 [26] と mT5 [34] のマルチタスク学習を利用して，アイヌ語の方言とドメインクラス毎にタスク接頭辞 (Task prefix) を付し，単一の方言・ドメインのみのクラス，並びに複数の方言・ドメインを混ぜた混合コーパスを用いて NMT モデルを学習した [11]．混合コーパスを用いたモデルは単一の方言・ドメインのクラスで学習したモデルよりも翻訳精度が高かったと報告しており，特に会話体データの少ない千歳方言や幌別方言と日本語との翻訳では非常に高い BLEU スコアを達成した (アイヌ語千歳方言→日本語: 70.97，アイヌ語幌別方言→日本語: 89.19，日本語→アイヌ語千歳方言: 70.96，日本語→アイヌ語幌別方言: 80.89) ．精度が向上した理由として，特に日本語からアイヌ語の各方言への翻訳の際に，対象となる方言をタスク接頭辞によって指定したことで，コーパス内の方言の違いによる曖昧さを解消させることができたことで効率的な学習が行われたと考察している．

2.3　低言語資源の言語を対象とした機械翻訳

利用可能な言語資源，特に機械翻訳の場合にはパラレルコーパスの量が十分に存在しない言語は低言語資源の言語 (あるいは低資源言語) と呼ばれる．英語や中国語など多くの使用者が存在する言語に比べて，使用者が少ない言語は利用可能なパラレルコーパスが十分ではなく，精度の高い NMT モデルを学習できないこと

が多い．本節では言語資源が十分に存在しない言語の機械翻訳に取り組んだ先行研究について述べる．

2.3.1 反復学習を利用した機械翻訳・コーパス拡張

反復学習によるコーパス拡張や翻訳モデルの精度向上に関する先行研究について述べる．

同語族言語間の機械翻訳 Przystupa らは，言語の類似性を考慮した NMT モデルの性能向上の研究として，低言語資源環境下におけるスペイン語・ポルトガル語間，チェコ語・ポーランド語間，ヒンディー語・ネパール語間のパラレルコーパスを逆翻訳によって拡張する手法を提案した [25]．初期のパラレルコーパスから注意機構付き LSTM モデル，Transformer モデルによる逆翻訳モデルを学習し，それを用いて目標言語の文を原言語の文に翻訳することで新たな翻訳ペアを獲得した．さらに，NMT モデルの学習と逆翻訳による翻訳ペアの生成を 3 回繰り返すことでパラレルコーパスを漸進的に増強した．ネパール語からヒンディー語への翻訳モデル以外，3 回反復したうち 3 回目の反復 (Synth3) で最も高い翻訳精度が得られた．ヒンディー語・ネパール語間の翻訳では，1 回目の反復 (Synth1) の時点でヒンディー語→ネパール語で 5 倍，ネパール語→ヒンディー語で 10 倍のアップサンプリングを実施したところ，ネパール語→ヒンディー語の翻訳モデルは精度が向上したが，ヒンディー語→ネパール語の翻訳モデルの顕著な精度向上は認められなかったと報告している．この要因はヒンディー語文とネパール語文の比率が合成テキストデータセット内で変化したことにあると考察しており，データセット内の両言語文の量を変化させない工夫が必要であると論じている．

ブリブリ語・スペイン語間の機械翻訳 Feldman らは，反復的逆翻訳の手法を用いて，コスタリカの先住民族言語の一つであるブリブリ語とスペイン語のニューラル機械翻訳モデルの翻訳精度を向上させた [8]．この研究では，少量のパラレルコーパスを用いて，ブリブリ語→スペイン語 NMT モデルを学習し，そのモデルの出力をブリブリ語文の対訳候補として初期パラレルコーパスに追加し，その拡張後パラレルコーパスを用いてスペイン語→ブリブリ語 NMT モデルを再学習した．パラレルコーパスから 1/4 (1,480 ペア)，1/2 (2,961 ペア)，全部 (5,923 ペア) をそれぞれ取り出し，これらを初期データセットとしてベースラインの Transformer モデルを訓練したところ，10 回の試行の BLEU スコアの平均がそれぞれ 6.3 ± 2.0, 11.2 ± 1.9，16.9 ± 1.7，最大でそれぞれ 9.9，14.9，19.8 のスコアであった．反復的逆翻訳によってこれらのベースモデルを再学習したところ，1 度目の反復処理では，5 回の試行の平均でそれぞれ -2.5 ± 0.8，1.0 ± 0.9，-1.9 ± 0.5，最大スコアはそれぞれ -1.9，2.1，-1.2 変化したと報告している．2 回目の反復の時にはどのデータサイズでも 1 回目よりも平均値が下がり，最大改善値についても 5,923 ペア

を用いてベースモデルを作成した再学習モデルを除き，BLEU スコアは改善しなかったと報告している．

英語 ↔ ベトナム語間の双方向機械翻訳モデル　Bui らは，対訳コーパスを用いた双方向翻訳モデルの反復的な更新手法を提案した [30]．IWSLT2015 の英語・ベトナム語翻訳タスクの対訳コーパス (13 万文) を使用した双方向翻訳の実験において，初期コーパスのみで学習した双方向翻訳モデルと比べ BLEU スコアが英語→ベトナム語方向で+2.61，ベトナム語→英語方向で+3.05 向上したと報告している．この研究は元の対訳コーパスのみを学習データとしており，言語資源の少ない状況においても双方向翻訳モデルの両方向の翻訳性能を有意に改善できると結論づけている．

2.3.2　反復学習を利用しない機械翻訳・コーパス拡張

反復学習を利用しない方法によってコーパス拡張や翻訳モデルの精度向上を実現した先行研究について述べる．

ピボット言語の自動選択アルゴリズムを用いたニューラル機械翻訳　ピボット機械翻訳は，原言語→目標言語の翻訳を実施する代わりに，比較的パラレルコーパスが多く翻訳精度も高い中間言語を利用し，原言語→中間言語→目標言語の順で翻訳を実施する手法である [33]．Leng らは，中間言語を用いたピボット機械翻訳の手法を発展させ，言語系統の異なる原言語から目標言語への翻訳タスクにおいて，ピボット翻訳で利用する中間言語を適切に自動選択する手法 LTR (Learn to route) を提案した [35]．20 言語，合計 294 の言語ペアを対象に評価実験が行われた．デンマーク語からガリシア語への翻訳の例では LTR によるピボット翻訳を行わないベースモデルと比較して BLEU スコアが最大 5.58 ポイント向上した．また，系統の離れた言語間の翻訳モデルを直接訓練するより，訓練なしピボット翻訳を用いる方が高い精度が得られると報告している．

琉日翻訳　久高は，沖縄語 (論文中では「琉球方言」と呼ばれている) から日本語への統計的機械翻訳 (Statistical Machine Translation: SMT) の精度向上を目的とし，対訳コーパスから自動生成した新しい対訳文を対訳コーパスに追加しコーパス拡張した上で，SMT モデルを学習した [17]．拡張後コーパスを用いた SMT モデルは拡張なしのコーパスを用いたモデルと比べ BLEU が最大 1.24 ポイント，RIBES が最大 2.54 ポイント向上したと報告している．また，対訳コーパスが多ければ多いほど精度が向上するわけではなく，十分な語彙や文脈を含む対訳コーパスが得られない場合があるため，合成パラレルコーパスに多様性を持たせつつ，その量を適切に調整する必要があると考察している．

高地ソルブ語・ドイツ語間の機械翻訳　Berckmann と Hiziroglu は，ドイツ連邦共和国領内の少数民族言語である高地ソルブ語とドイツ語を対象とした NMT モデルを転移学習する手法を提案した [1]．彼らは Transformer アーキテクチャの Decoder のみを NMT モデルとして採用した．それぞれの対象言語と同語派のチェコ語・英語パラレルコーパス (60M) を用いて NMT モデルを事前学習し，次に高地ソルブ語・ドイツ語パラレルコーパス (60K) を用いて NMT モデルをファインチューニングした．少量の高地ソルブ語・ドイツ語パラレルコーパスのみを用いて学習した NMT モデルと比較すると，BLEU スコアがドイツ語→高地ソルブ語方向の翻訳で+17.2 ポイント，高地ソルブ語→ドイツ語方向の翻訳で+15.6 ポイント向上し，低言語資源環境における翻訳精度の向上には類似する言語を使用した NMT モデルの事前学習は有効であったと報告している．

2.4　本研究の特色

本研究では，低言語資源の言語であるアイヌ語から日本語へのニューラル機械翻訳に反復的逆翻訳 [9] を適用し，翻訳モデルの精度向上を図る．さらに，目標言語である日本語文を英語文に翻訳し，それを再度日本語文に翻訳することで，パラレルコーパスをより大きく拡張する手法を提案する．本提案手法で学習された NMT モデルと，反復なしで学習したベースモデル，先行研究の反復的逆翻訳によって学習された NMT モデルについて，翻訳精度を測定し，比較する．

本研究は低言語資源環境下の翻訳タスクに対して逆翻訳を適用する点で，Feldman らや Przystupa らの研究と近いが，本研究の提案手法では目標言語文の拡張において原言語を用いた拡張のみではなく，第三の言語を用いて拡張する処理を追加している点が異なる．また，北海道アイヌ語の機械翻訳に反復的逆翻訳によるコーパス拡張を適用した研究は著者の知る限り本研究が初めてである．

第3章　アイヌ語・日本語パラレルコーパスの整備

本章では本研究で使用するアイヌ語・日本語パラレルコーパスについて述べる．3.1 節ではパラレルコーパスとして利用する文献について述べる．3.2 節ではアイヌ語の表記の統一について述べる．3.3 節ではパラレルコーパスの形式と例とともに説明する．

3.1　利用文献

2024 年 10 月現在，一般に公開されており，かつ形式が統一されたパラレルコーパスが存在しないことから，いくつかの文献からアイヌ語・日本語のテキストを抽出し，人手で形式を統一してアイヌ語・日本語パラレルコーパスを構築する．具体的には，表 3.1 に記載した文献からテキストを抽出し，綴りやデータ形式を統一したパラレルコーパスを整備した．なお，本研究で利用するアイヌ語文献はすべて北海道アイヌ語 [[hokk1243]] の文献のみで，樺太アイヌ語 [[tara1248]] [[sakh1245]] や千島アイヌ語 [[kuri1271]] の文献は含んでいない．[1]

アイヌ語・日本語のコーパスは文献 [4, 15, 16, 20, 31] のように整備されインターネット上に公開されているものがあるが，文献 [2] のように文献が電子媒体で閲覧可能になっているのみで，コーパスとして利用できるように整備されていないものある．電子データで公開されていない文献については，OCR によってテキストデータを取得し，本研究で使用するパラレルコーパスに含めた．また，管理機関の違いからそれぞれのコーパスの形式が異なるため，一つのコーパスに統合するにはデータ形式の調整が必要である．調整の詳細は 3.2 節，3.3 節で述べる．

利用データに関して，アイヌ語文献の多くはブガエワ [3] にあるように口承文芸が多く，よって本研究で使用するコーパスも主となるドメインは口承文芸であるが，文献 [10, 12, 16] のように会話体のテキストも含まれる．

また，2024 年 10 月現在，アイヌ語はいわゆる「共通語」や「標準語」と呼ばれる言語変種が整備されておらず，かつ地域方言間での差異も見られる．しかし，コーパス量を確保することを優先し，地域方言を区別せずにパラレルコーパスに

[1] [[ ]] 内は Glottolog コード: `https://glottolog.org/resource/languoid/id/ainu1252` (最終閲覧日: 2025 年 1 月 18 日)

表 3.1: アイヌ語・日本語パラレルコーパスの出典となる文献の一覧

管理番号	文献	地域	資料数
C00001	アイヌ語会話辞典コーパス [16]	-	1
C00002	アイヌ語口承文芸コーパス [20]	千歳	38
C00003	アイヌ語鵡川方言 日本語 – アイヌ語辞典 [4]	鵡川	1
C00004	二風谷アイヌ民族博物館 [1][2]	沙流	98
C00005	アイヌ語アーカイブ [15]	千歳，沙流，静内	201
C00006	AA 研アイヌ語資料 [31]	千歳，静内	31
D00001	萱野辞典 [13]	沙流	1
T00001	アコロ イタク [10]	千歳，沙流，旭川	1
T00002	アイヌ民譚集 [5]	幌別	15
T00003	アイヌ神謡集 [21]	幌別	13
T00004	「1991 年 11 月 10 日 千歳市蘭越生活館における アイヌ語会話ビデオ」内会話 [3][12]	千歳	1

[1] 2025 年 1 月現在，国立アイヌ民族博物館アイヌ語アーカイブ [15] に移管
[2] うち，アコロ イタクと重複する文章は除く

含める．この結果，本研究のパラレルコーパスには幌別，沙流，千歳，鵡川，静内を中心に複数地域の方言の文が含まれている．

3.2 表記形式の統一

前節に挙げたアイヌ語の複数の文献では表記が統一されていないため，これを同じ表記に変換する作業を行う．以下，項目別に表記形式の統一作業の詳細について述べる．

アイヌ語の綴りの統一 アイヌ語は，2024 年現在，正書法が定義されておらず，文献によって綴りに揺れがある．本論文で作成したアイヌ語・日本語パラレルコーパスにおけるアイヌ語の表記はすべて文献 [10] のローマ字表記法に従った．文献 [10] 以前に書かれた，文献 [5, 5, 21] のアイヌ語表記については，綴りを原文献から大きく変更した．アイヌ語がアイヌ語カナで表記されている文献に関しては文献 [10] に倣い，ローマ字表記に変換した．

次の例 (1) はアイヌ民譚集 [5] における散文説話 (*ain.* uwepeker)『パナンペ銀の子犬を授かる』の一文である．1 行目は原文表記，2 行目は本コーパスにおける綴り統一後の表記を表している．3 行目のグロス表記 (それぞれの語の英訳) は参考情報であり，文献 [16, 20] を参考に付与した．4 行目は日本語訳である．

(1) アイヌ民譚集 [5]

chip shik kane kamui-chep kusa wa ek aike
cip sik kane kamuy cep kusa wa ek ayke
boat luggage doing.so salmon take.by.boat and come then

「舟いっぱいに鮭を運んで来ると」

アイヌ語の形態素境界の統一 アイヌ語では動詞を使用する際にその主語や目的語に対応する人称接辞を用いた表示が必須である．本コーパスでは，動詞や名詞に人称接辞が接合している場合，文献 [10] に従い，形態素境界に等号 (=) を挿入している．人称接辞と接合先の語に形態素境界としての等号が付されていない場合には等号を新たに付与し，等号の代わりに黒点 (・) などが利用されている場合は等号に置き換えた．また，本コーパスでは，人称接辞は付加される先の形態素と組み合わせた形ではなく，それぞれを別々の形態素として取り扱うこととする．このルールにより，k(u)=，e=，eci=，es=，c(i)=，a=，an=，aci=[2]，en=，un=，i=，=as，=an の各人称接辞はその接合対象である動詞や名詞と表記上切り離され，語境界である半角空白によって独立した形態素として記載されている．なお，人称接辞と接合先対象の語を切り離す表記はアイヌ語学で一般的に用いられる表記ではないことは注意されたい．

次の例 (2) はアイヌ語会話辞典コーパス [5] に掲載の沙流方言の例文である．文中の kus (*jpn.* 通る) と arpa (*jpn.* 行く) はそれぞれ動詞で，これに 1 人称単数主格接辞である k(u)=が接合している．沙流方言を含む日高西部ならびに千歳方言では，ku= (1 人称単数主格接辞) が母音から始まる動詞と接合する場合，接辞内の母音が脱落し，k=という形式になる [22]．そのため，例 (2) においても，母音から始まる arpa に接合された結果，k=arpa という形式になっている．原文では ku=kus と k=arpa は一体表記されているが，本コーパスでは ku=と kus，k=と arpa を独立した形態素として記載している．

(2) アイヌ語会話辞典コーパス [16]

sirteksam ku=kus wa k=arpa.
sirteksam ku= kus wa k= arpa
appearance-hand-near 1SG.A=pass and 1SG.S=go.SG

「私は海岸沿いを通って行く.」

人称接辞以外の接辞や語に関して，アイヌ語では充当接頭辞 (e-，o-，ko-等) や使役形化接尾辞 (-re 等)，反復接尾辞 (-pa) など多くの接辞が用いられる

[2]千歳方言で稀に用いられる a=eci= (4 人称主格・2 人称複数目的格接辞) の縮約形 [2, 3, 22]

[22, 15]．文献によって語幹と接辞の間にハイフン (-) が挿入されているものや，また例 (3) に示すように，名詞句と格助詞の間に語境界としてハイフンが含まれる文については，綴りの揺らぎを極力減らすため，形態素間のハイフンを半角空白に置き換えた．

(3) アイヌ民譚集 [5]

shineanto-ta Panampe pish-ta san, inkar aike
sineanto ta Panampe pis ta san, inkar ayke
one.day at Panampe beach at front.place=INTR.SG look then

「或日，パナンペが浜へ出て見ると」

例 (4) に示すように，存在格を表す or ta を otta と記載している場合など，連続する 2 音節間で音素交替がある場合，当該形態素を辞書形の綴りに戻し音素交替が起こった箇所にはアンダーバー (_) を付した．

(4) アイヌ神謡集 [21]

otta unante.
or_ ta un= ante.
place at 1PL.O= exist.SG-cause

「私をそこへ置きました.」

また，特にカナ表記されていた文献 [12] のテキストをローマ字変換する際に，アンシェヌマンにより，2 つ以上の形態素が 1 つの語に結合されて表記されている箇所については，各種辞典を参考に語ごとに半角空白を語境界として単語分割を行った．例 (5) では,「クトテク」を「ku=」と「totek」の 2 つの単語に分割している．

(5) 「1991 年 11 月 10 日 千歳市蘭越生活館におけるアイヌ語会話ビデオ」内会話 [12]

クトテク ワ カン マ.
ku= totek wa k= an _wa.
1SG.S= be.healthy and 1SG.S= exist.SG FIN

「(私は) 元気でいるよ」

行区切りの統一 利用した文献の中に，書面もしくはディスプレイの都合上文中で改行が入った文献がある．これらの文についてはピリオドや句点までを一文として一行にまとめた．対応する日本語文もアイヌ語に対となるように行区切りを調整した．

翻訳対象外箇所への削除マーカー付与 各文献において，kamuyyukar (*jpn.* 神謡) に含まれる sakehe (*jpn.* 折返) やそれに準ずる繰り返しのフレーズ，録音時のコメントや註釈，フィラー，言い間違いなど機械翻訳を行うにあたって不要な箇所は翻訳対象外として波括弧 ({}) で括った．

3.3 コーパスの形式

各文献から抽出したアイヌ語と日本語の文の組を JSON 形式で保存し，文献データセットを作成した．図 3.1 はこの文献データセットの一例である．code は対訳ペアの識別 ID，title_ain は物語のアイヌ語のタイトル，title_jpn は日本語のタイトル，1 は日本語の文 (jpn)，アイヌ語の文 (ain)，英語の文 (eng) から構成される対訳ペアを表す．前節で言及した sakehe が含まれる文献の場合は，該当のフレーズを sakehe キーに代入し，アイヌ語文中の該当箇所は V に置き換えた．

次にそれぞれの文献データセットの先頭からアイヌ語と日本語の文を抽出し，ID，アイヌ語文，日本語文をタブ区切り形式のテキストファイルに変換した．ID はテキストファイルに追加した順番で自動採番した．これに加えて，前節で言及した波括弧で括られた翻訳対象外箇所を削除した．この手順で作成されたテキストファイルを本論文における初期のパラレルコーパスとする．図 3.2 は本パラレルコーパスの一部を抜粋したものである．


```
{
    "code": "K7807152KY",
    "title_ain": "",
    "title_jpn": "カムイユカラ：火の女神が地の果ての魔神と戦った",
    "sakehe_v1_latin":"apemerumeru koyankoyan mat",
    "sakehe_v2_latin":"ateyatenna",
    "1": {
        "jpn": "夜も昼も針仕事ばかり私はしていた。",
        "ain": "{V1}{V2}{u} kunne hene {V1}{V2}{u} tokap hene {V1}{V2} kemeyki patek
{V1} ci= ki pa ki na.",
        "eng": "During the nights, during the days I only [did] needle work.I did it."
    },
```


図 3.1: 文献データセットにおける対訳ペアの例

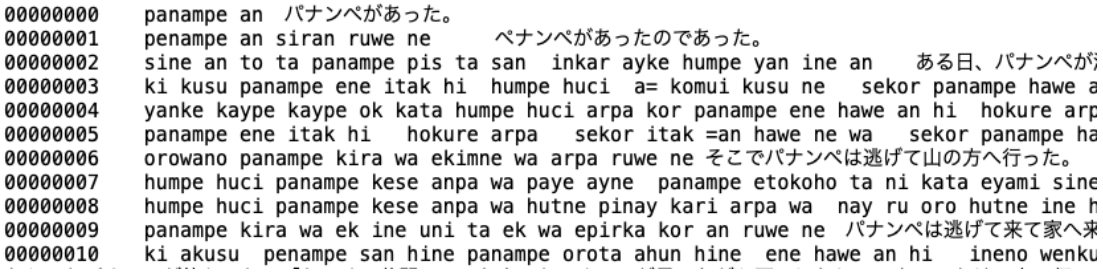

図 3.2: アイヌ語・日本語パラレルコーパスの抜粋

第4章　提案手法

本章では本研究の提案手法について述べる．

4.1　概要

提案手法の概要を図 4.1 に示す．原言語 (本研究ではアイヌ語) は src，目標言語 (本研究では日本語) は tgt と記す．基本的には，翻訳モデルの学習とそれを用いたコーパス拡張を繰り返す反復学習を行い，パラレルコーパスの規模を漸進的に拡大することで翻訳モデルの品質を高める．

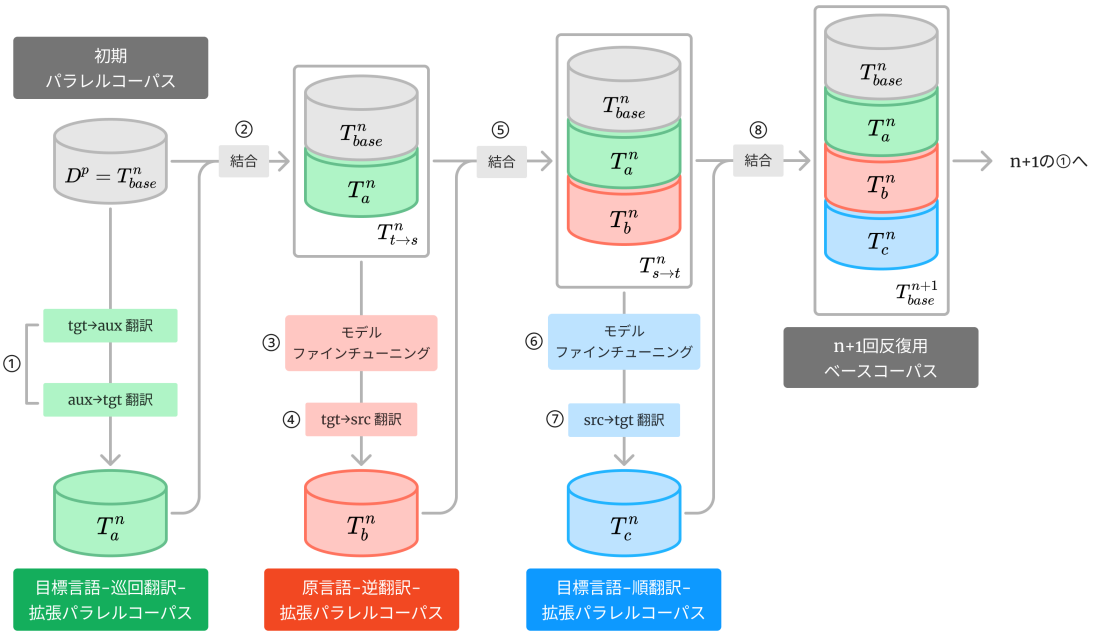

図 4.1: 提案手法の概要

反復学習の各ステップにおいて，コーパスの拡張は以下に示す ⟨1⟩，⟨2⟩，⟨3⟩ の 3 段階に分けて行う．以降の説明において，T^n_{base} は n 回目の反復学習の開始時点でのパラレルコーパスを表す．反復回数が 1 のとき，T^1_{base} は初期のアイヌ語・日本語のパラレルコーパス (D^p と記す) と等しい．

⟨1⟩ T^n_{base} から「目標言語-巡回翻訳-拡張パラレルコーパス」T^n_a を得る．

⟨2⟩ T^n_{base} と T^n_a から「原言語-逆翻訳-拡張パラレルコーパス」T^n_b を得る．

⟨3⟩ T^n_{base}，T^n_a，T^n_b から「目標言語-順翻訳-拡張パラレルコーパス」T^n_c を得る．

上記の手続きを経て得られる $T^n_{base} \cup T^n_a \cup T^n_b \cup T^n_c$ が次の反復ステップで用いる最初のパラレルコーパス T^{n+1}_{base} となる．

各段階のコーパス拡張は以下のように行う．

1. 目標言語-巡回翻訳-拡張パラレルコーパス (T^n_a) の構築

   「巡回翻訳」によって「目標言語」の文を新たに生成することでパラレルコーパスを拡張する．元のパラレルコーパスにおける原言語の文 s と目標言語の文 t の組を (s,t) とおく．図 4.1 の①に示す巡回翻訳，すなわち t を補助言語 aux に翻訳し，さらにその文を目標言語 tgt に翻訳することで，新しい目標言語の文 $\hat{t}$ を得る．そして，$(s,\hat{t})$ を新しい翻訳ペアとして T^n_a を構築する．

2. 原言語-逆翻訳-拡張パラレルコーパス (T^n_b) の構築

   「逆翻訳」によって「原言語」の文を新たに生成することでパラレルコーパスを拡張する．図 4.1 の③に示すように，T^n_{base} と T^n_a を結合したパラレルコーパス ($T^n_{t\to s}$ と記す) を学習データとし，$tgt \to src$ の逆翻訳モデルをファインチューニングする．以下，この逆翻訳モデルを $M^{Cn}_{t\to s}$ と記す.[1]④に示すように，翻訳モデル $M^{Cn}_{t\to s}$ を用いて，(s,t) における t を逆翻訳して新しい原言語の文 $\hat{s}$ を得て，$(\hat{s},t)$ を新しい翻訳ペアとして T^n_b を構築する．

3. 目標言語-順翻訳-拡張パラレルコーパス (T^n_c) の構築

   「順翻訳」によって「目標言語」の文を新たに生成することでパラレルコーパスを拡張する．図 4.1 の⑥に示すように，T^n_{base}，T^n_a，T^n_b を結合したパラレルコーパス ($T^n_{s\to t}$ と記す) を学習データとし，$src \to tgt$ の (順方向) 翻訳モデルをファインチューニングする．以下，この逆翻訳モデルを $M^{Cn}_{s\to t}$ と記す．⑦に示すように，翻訳モデル $M^{Cn}_{s\to t}$ を用いて，(s,t) における s を翻訳して新しい目標言語の文 $\hat{t}$ を得て，$(s,\hat{t})$ を新しい翻訳ペアとして T^n_c を構築する．

以上の反復学習の目的は原言語から目標言語への翻訳モデルを学習することだが，それは操作⑥にて得られる翻訳モデル $M^{Cn}_{s\to t}$ である．コーパス拡張により，初期のパラレルコーパスよりも規模の大きいコーパスを用いてモデルを学習することができるため，翻訳品質の向上が期待できる．また，コーパス拡張と翻訳モデルの学習を交互に反復することで，翻訳モデルの品質も漸進的に改善されることが期待される．

[1] 上付き文字の C は提案手法「反復的巡回翻訳」の巡回 (Cyclic)，n は反復回数，下付き文字は翻訳方向 (t, s は目標言語，原言語) を表す．

本研究では上記の提案手法を「反復的巡回・逆・順翻訳」と呼ぶ．また，これ以降本論文では特筆がない限り，反復的巡回・逆・順翻訳を「反復的巡回翻訳」と略記する．

以下の節では，それぞれの処理の詳細を述べる．4.2 節では機械翻訳の基準モデルについて説明する．4.3 節では補助言語を介した巡回翻訳による目標言語文の拡張 (図 4.1 の①，②) について述べる．4.4 節では逆翻訳を利用した原言語文の拡張 (図 4.1 の③，④) と順翻訳を利用した目標言語文の拡張 (図 4.1 の⑥，⑦) について述べる．4.5 節では提案手法の反復学習全体を詳述する．

4.2 基準モデル

本節では，ベースとする NMT モデルについて述べる．本研究では，利用可能なアイヌ語・日本語パラレルコーパスの量が多くないことから，Transformer 等の深層学習モデルを最初から学習することは行わず，パラレルコーパス中の翻訳ペアを使用して事前学習済み NMT モデルをファインチューニングすることで翻訳モデルを学習する．事前学習済み NMT モデルとして，Meta 社によって公開されている nllb-200-distilled-600M[2]を用いる．

初期のパラレルコーパス D^p を使用して，原言語から目標言語の翻訳を行う NMT モデル $M^{base}_{s\to t}$ をファインチューニングにより構築する．以下，これを基準モデルと呼ぶ．評価実験では，基準モデルをベースラインとし，本研究で提案する反復学習によってどれだけ翻訳精度を向上させることができるかを評価する．

4.3 補助言語を介した目標言語文の拡張

本節では，低言語資源環境下での有効的なコーパス拡張方法である補助言語 aux を介した目標言語 tgt のデータ拡張の手順について記述する．目標言語文の補助言語 aux を介した巡回翻訳を経て新たな翻訳候補文を獲得することで，パラレルコーパスの拡張と翻訳モデルの品質向上を狙う．

翻訳対象となる目標言語文は反復開始時点で保持しているパラレルコーパス T^n_{base} の目標言語文である．初回反復時 (n=1) の時は初期コーパス D^p に含まれる目標言語文が対象となる．サードパーティーの機械翻訳システムを利用し，パラレルコーパスにおける翻訳ペアの目標言語の文 t を補助言語 aux に一度翻訳し，次にこの出力文を再度同システムを利用し，目標言語文に再翻訳する．この処理で得られた文 $\hat{t}_a$ を元の原言語文 s の新たな翻訳結果として対訳ペアを作成し，目標言語-巡回翻訳-拡張パラレルコーパス T^n_a に格納する．上記の処理は図 4.1 の①に該当する．

[2]https://huggingface.co/facebook/nllb-200-distilled-600M

本研究では，補助言語 aux として英語を選択する．また，本研究では目標言語から補助言語，ならびにその逆方向への機械翻訳精度の向上は目標としていないため，新たにこのためにモデルの訓練やファインチューニングは実施せず，翻訳精度がある程度高く容易に利用が可能な Google 翻訳を利用した．また巡回翻訳の実装には，Python 向けの機械翻訳ライブラリ googletrans[3]を採用した．

なお，2 回目の反復 (n >1) 以降における翻訳対象文は，その前回の反復までに一度も補助言語を介した翻訳をしておらず，かつ再々度の巡回翻訳にならない目標言語文とする．例えば，n=2 のときの翻訳対象は，前回反復 (n=1) 時に得られた原言語→目標言語翻訳モデル $M_{s\to t}^{C1}$ の出力文 $\hat{t}$ (ただし，$\hat{t} \in T_c^1$) のみとなる．これを補助言語を介して巡回翻訳を行い，新たな翻訳候補 $\hat{t}_a^2$ を得る．

上記の手続きで得られた拡張コーパス T_a^n を元コーパス T_{base}^n へ追加し，目標言語→原言語方向の翻訳モデルのファインチューニング用パラレルコーパス $T_{t\to s}^n$ を作成する．これは図 4.1 の②に該当する．

巡回翻訳を用いて目標言語の拡張を行う動機は以下の通りである．高資源言語については既に精度の高い機械翻訳システムが利用可能であるため，これを利用した巡回翻訳によって日本語文を別の文に言い換えることで，高品質のパラレルコーパスを新たに獲得できる．補助言語として英語を選択したのは，既に述べたように，既存の日本語-英語間の翻訳システムの精度が高いためである．なお，原理的には原言語に巡回翻訳を適用してコーパスを拡張することも可能だが，原言語であるアイヌ語は低言語資源の言語であり，精度の高い翻訳システムは利用できないため，本研究では実施しない．

4.4 逆翻訳・順翻訳による原言語文・目標言語文の拡張

本節では逆翻訳による原言語文の拡張手順，ならびに順翻訳による目標言語文の拡張手順について説明する．

逆翻訳による原言語文の拡張は以下の手続きで行う．まず，前節で述べた手続きで得られるパラレルコーパス $T_{t\to s}^n$ を用いて事前学習済み機械翻訳モデルをファインチューニングし，目標言語→原言語方向の機械翻訳モデル $M_{t\to s}^{Cn}$ を学習する (図 4.1 の③)．このモデルを用いて，パラレルコーパス $T_{t\to s}^n$ の目標言語文を原言語へ翻訳し，この出力文と元の目標言語文を組み合わせて新たな対訳ペアを作成し，原言語-逆翻訳-拡張パラレルコーパス T_b^n に格納する (図 4.1 の④)．このコーパス T_b^n を目標言語→原言語翻訳モデルのファインチューニング用パラレルコーパス $T_{t\to s}^n$ へ追加し，原言語→目標言語翻訳モデルのファインチューニング用パラレルコーパス $T_{s\to t}^n$ を作成する (図 4.1 の⑤)．

次に，順翻訳による目標語文の拡張を以下の手続きで行う．作成した拡張コーパス $T_{s\to t}^n$ を使用して新たに機械翻訳モデルをファインチューニングし，原言語→

[3] https://pypi.org/project/googletrans/

目標言語方向の機械翻訳モデル $M_{s\to t}^{Cn}$ を生成する (図 4.1 の⑥)．このモデルを用いて，パラレルコーパス $T_{s\to t}^{n}$ の原言語文を目標言語へ翻訳し，この出力文と元の原言語文を組み合わせて新たな対訳ペアを作成し，目標言語-順翻訳-拡張パラレルコーパス T_c^n に格納する (図 4.1 の⑦)．このコーパス T_c^n を原言語→目標言語翻訳モデルのファインチューニング用パラレルコーパス $T_{s\to t}^{n}$ に追加し，次の n+1 回目の反復開始時点で利用するパラレルコーパス T_{base}^{n+1} を作成する (図 4.1 の⑧)．

4.5 パラレルコーパスの反復拡張・翻訳モデルの反復学習のまとめ

提案手法の擬似コードを Algorithm 2 に示す．T は翻訳モデルの学習に用いるデータセット，すなわち原言語と目標言語の文の組 (s,t) の集合からなるパラレルコーパスを表す．一方，M は翻訳モデルを表す．データセット名やモデル名の下付き文字について，$s\to t$ は「アイヌ語→日本語方向の翻訳」，$t\to s$ は「日本語→アイヌ語方向の翻訳」をそれぞれ表す．n は反復学習におけるステップ数を，N は反復回数の上限を表す．

Algorithm 2 の擬似コードと図 4.1 の対応は以下の通りである．8 行目は補助言語を介した巡回翻訳による目標言語文生成によるコーパス拡張であり，図 4.1 の①に該当する．9 行目はパラレルコーパスの結合であり，図 4.1 の②に該当する．10 行目は逆方向翻訳モデル $M_{t\to s}^{Cn}$ の学習であり，図 4.1 の③に該当する．11 行目は逆方向翻訳モデルによる原言語文生成によるコーパス拡張であり，図 4.1 の④に該当する．12 行目はパラレルコーパスの結合であり，図 4.1 の⑤に該当する．13 行目は順方向翻訳モデル $M_{s\to t}^{Cn}$ の学習であり，図 4.1 の⑥に該当する．14 行目は順方向翻訳モデルによる目標言語文生成によるコーパス拡張であり，図 4.1 の⑦に該当する．15 行目はパラレルコーパスの結合であり，図 4.1 の⑧に該当する．

提案手法におけるコーパス拡張処理の詳細を図 4.2 に示す．この図では，どの文を言い換え・翻訳することで新しい翻訳ペア (拡張パラレルコーパス) を獲得するか，新しく構築したパラレルコーパスと既に構築済みのパラレルコーパスをどのように結合してパラレルコーパスを拡張していくかを明示している．

T_a^n，T_b^n，T_c^n $(n=1,2)$ は，これまで説明した「目標言語-巡回翻訳-拡張パラレルコーパス」,「原言語-逆翻訳-拡張パラレルコーパス」,「目標言語-順翻訳-拡張パラレルコーパス」に該当する．これらは以下の手続きで構築される．

- T_a^n は，補助言語を介した巡回翻訳によって，日本語文 Jpn_{org} ($n=1$ のとき) もしくは Jpn_{trl}^{n-1} ($n=2$ のとき) を別の日本語文 Jpn_{aux}^{n} に言い換えることによって得る．

- T_b^n は，パラレルコーパス $T_{t\to s}^{n}$ における日本語文 (図 4.3 の緑で示した文) を逆方向翻訳モデルによってアイヌ語文 Ain_{trl}^{n} に翻訳することで構築する．こ

Algorithm 2 提案手法

Input: original parallel corpus D^p

Let $T^1_{base} := D^p$
for $n = 1$ to N **do**
  if $n == 1$ **then**
    Let $T^n_t := T^n_{base}$
  else
    Let $T^n_t := T^{n-1}_c$
  end if
  Create $T^n_a := \{(s, \hat{t}_a)\}$ through $t \xrightarrow{M_{t\to a}} a \xrightarrow{M_{a\to t}} \hat{t}_a$, for $(s,t) \in T^n_t$
  Let $T^n_{t\to s} := T^n_{base} \cup T^n_a$
  Fine-tune target-to-source NMT model $M^{Cn}_{t\to s}$ using $T^n_{t\to s}$
  Let $T^n_b := \{(\hat{s}, t)\}$ where $t \xrightarrow{M^{Cn}_{t\to s}} \hat{s}$, for $(s,t) \in T^n_{t\to s}$
  Let $T^n_{s\to t} := T^n_{t\to s} \cup T^n_b$
  Fine-tune source-to-target NMT model $M^{Cn}_{s\to t}$ using $T^n_{s\to t}$
  Let $T^n_c := \{(s, \hat{t})\}$ where $s \xrightarrow{M^{Cn}_{s\to t}} \hat{t}$, for $(s,t) \in T^n_{s\to t}$
  Let $T^{n+1}_{base} := T^n_{s\to t} \cup T^n_c$
end for

Output: NMT models $M^{Cn}_{s\to t}$ and $M^{Cn}_{t\to s}$

のときに用いる逆方向翻訳モデルは $T^n_{t\to s}$ を用いて学習されたモデルである．

- T^n_c は，パラレルコーパス $T^n_{s\to t}$ におけるアイヌ語文 (図 4.3 の青で示した文) を順方向翻訳モデルによって日本語文 Jpn^n_{trl} に翻訳することで構築する．このときに用いる順方向翻訳モデルは $T^n_{s\to t}$ を用いて学習されたモデルである．

提案手法は Hoang らによる反復的逆翻訳の手法 [9] を拡張したものである．その擬似コードを Algorithm 1 (2 章に掲載) に，コーパス拡張処理の詳細を図 4.3 に示す．既存の反復的逆翻訳と本研究で提案する反復的巡回翻訳の違いは，目標言語の巡回翻訳によるコーパス拡張を行うか否かである．すなわち，反復的逆翻訳では原言語への逆翻訳によるコーパス拡張と目標言語への順翻訳によるコーパス拡張を繰り返すが，反復的巡回翻訳ではこれに加えて目標言語の巡回翻訳によるコーパス拡張の処理が加わる．拡張パラレルコーパス T^1_a を構築する手続きに着目すると，図 4.3 に示す反復的逆翻訳では，逆翻訳モデルを学習する元となるパラレルコーパス $T^1_{t\to s}$ は初期のパラレルコーパスにおける翻訳ペア (Ain_{org}, Jpn_{org}) のみから構成されているのに対し，図 4.2 に示す反復的巡回翻訳では，$T^1_{t\to s}$ は (Ain_{org}, Jpn_{org}) と (Ain_{org}, Jpn^1_{aux}) から構成されている．

すなわち，逆翻訳モデルの学習に用いるパラレルコーパスが巡回翻訳によって増強されている．より大規模なパラレルコーパスを用いることにより，逆翻訳モ

デルならびにそれを用いて構築された拡張パラレルコーパスの品質を高めることができると期待できる．反復学習全体に着目すると，巡回翻訳によるコーパス拡張の処理を加えることで，反復的逆翻訳よりも最終的に得られるパラレルコーパスの量が多くなるため，それから学習する翻訳モデルの翻訳精度の向上が期待できる．

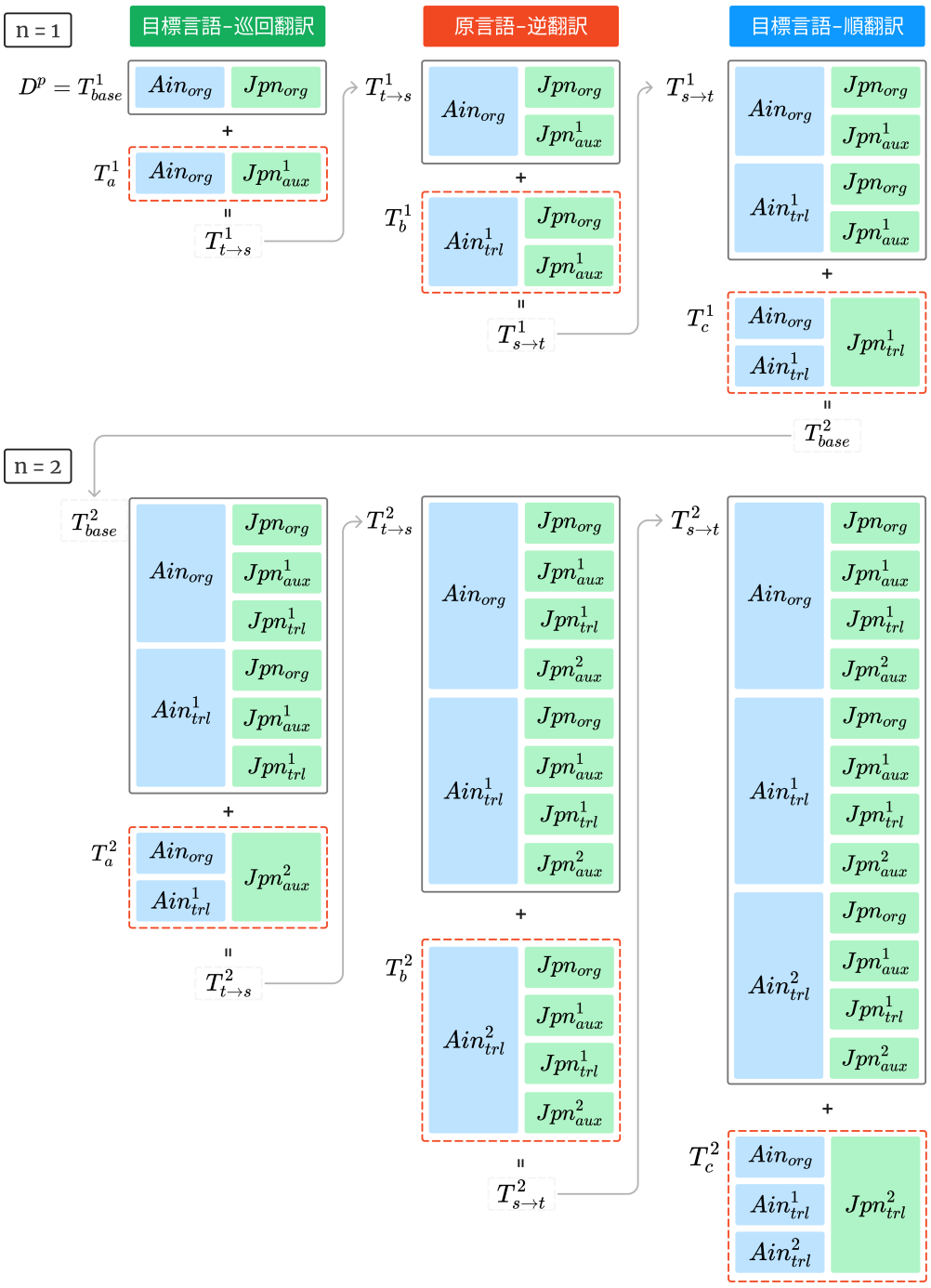

図 4.2: 反復的巡回翻訳におけるコーパス拡張処理の流れ

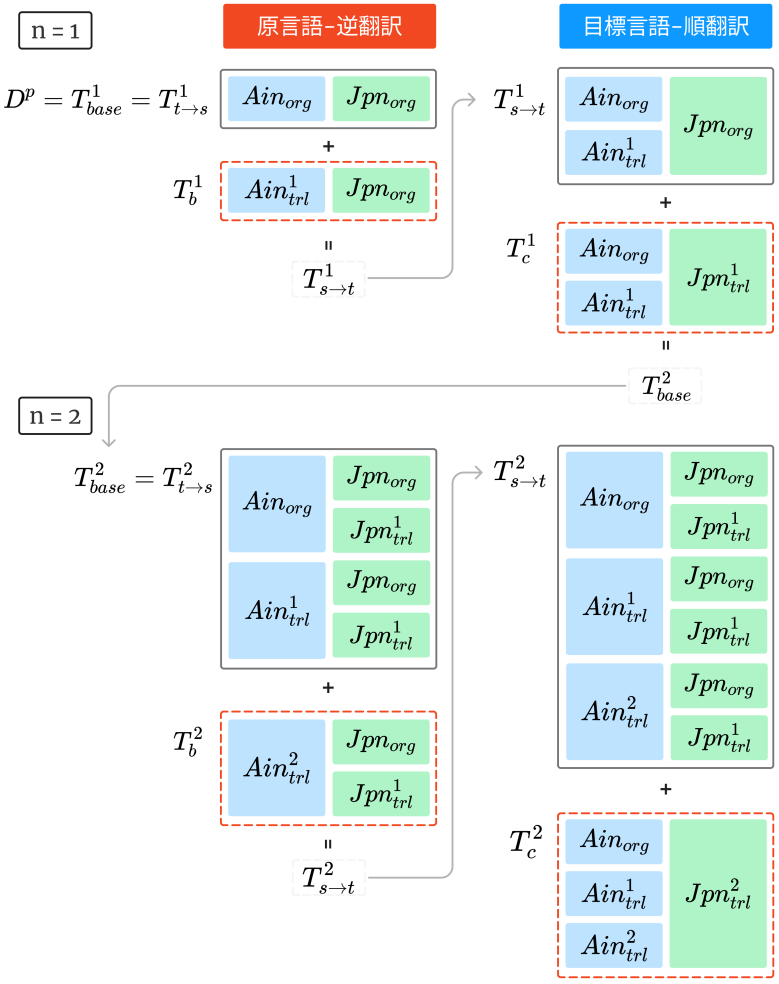

図 4.3: 反復的逆翻訳 [9] によるコーパス拡張処理の流れ

第5章　実験・評価

本章では，提案手法の評価実験について述べる．実験の使用データ，評価尺度，実験の結果，考察を順に説明する．

5.1　実験設定

5.1.1　使用データ・モデル

本実験では第3章で整備したアイヌ語・日本語パラレルコーパスを使用した．初期のコーパスのサイズ T^1_{base} を表5.1に示す．このコーパスを訓練データ $train$，検証データ val，テストデータ $test$ を8:1:1の割合で分割する．訓練データは翻訳モデルのファインチューニングによるパラメータ更新やコーパスの拡張に，検証データはファインチューニング中のモデルの精度評価に，テストデータはモデルの翻訳性能を測るために用いる．検証データ，テストデータは5.1.3項で述べる自動評価指標の算出のみで用い，モデルのファインチューニング時には利用しない．

表 5.1: 初期パラレルコーパスの翻訳ペア数

	train	val	test
T^1_{base}	18,669	2,334	2,334

データの前処理として，人称接辞と接合先である動詞，名詞，副詞との形態素境界に人称接辞マーカーとして用いられる等号 (=) を除いて，全ての記号を削除した．これにより，第3章で述べたコーパス整備時に残していた音素交替マーカー (_) もデータセットから除外された．アイヌ語では人称接辞が単独で存在せず，動詞や特定の名詞に義務的に付く性質上 [22]，人称接辞に接合している等号は文法上重要な意味合いを持つと仮定し，本実験においては除外対象としていない．

5.1.2 使用モデル

本研究では Transformer 等の深層学習モデルを最初から学習することは行わず，事前学習済み NMT モデルである nllb-200-distilled-600M[1] を使用する．トークナイザーはモデルのファインチューニング時に生成されたものを用いる．初期のパラレルコーパスのみをファインチューニングに用いるベースラインモデルに加え，提案手法である反復的巡回翻訳で得られたモデルと，既存手法である反復的逆翻訳 [9] で得られたモデルを比較する．反復回数の最大値を 2 と設定する．翻訳精度の評価対象となるアイヌ語→日本語翻訳モデルの一覧を以下に示す．なお，モデル名につく添え字について，上付き文字の C は提案手法の「巡回」 (Cyclic) ，B は既存手法の「逆」 (Back) ，base はベースモデル，数字は反復回数，下付き文字は翻訳方向 (s, t は原言語，目標言語) を表している．

ベースラインモデル $M_{s\to t}^{base}$**:** 拡張処理なしで初期パラレルコーパスのみを用いてファインチューニングを行い学習したモデル

反復的巡回翻訳による拡張モデル $M_{s\to t}^{C1}$**:** 提案手法を適用し，1 回目の反復で得られた拡張パラレルコーパスを用いてファインチューニングを行い学習したモデル

反復的巡回翻訳による拡張モデル $M_{s\to t}^{C2}$**:** 提案手法を適用し，2 回目の反復で得られた拡張パラレルコーパスを用いてファインチューニングを行い学習したモデル

反復的逆翻訳 [9] による拡張モデル $M_{s\to t}^{B1}$**:** 既存手法による拡張を実施し，1 回目の反復で得られた拡張パラレルコーパスを用いてファインチューニングを行い学習したモデル

反復的逆翻訳 [9] による拡張モデル $M_{s\to t}^{B2}$**:** 既存手法による拡張を実施し，2 回目の反復で得られた拡張パラレルコーパスを用いてファインチューニングを行い学習したモデル

5.1.3 評価方法

自動評価指標として，BLEU，chrF^{++} の 2 つを用いる．実験を通して獲得したアイヌ語から日本語方向の NMT モデルを用いて初期パラレルコーパスのテストデータのアイヌ語文を翻訳し，元の日本語訳との一致度をこれらの自動評価指標を用いて評価する．それぞれの評価指標は Python のライブラリである sacrebleu[14] を使用して算出し，小数点以下 2 桁に四捨五入した値を報告する．それぞれの評価指標について以下に述べる．

[1] https://huggingface.co/facebook/nllb-200-distilled-600M

BLEU BLEU (Bilingual Evaluation Understudy) は機械翻訳の出力文と参照訳との類似度を測る指標である [24]．BLEU は出力文の単語 N-gram が参照訳 (正解の訳) にどれだけ含まれているかを評価する．BLEU の値は 0%〜100% の値を取り，100%に近いほど参照訳と出力文が類似していることを示す．限られた長さのトークン列 (単語列) のみを単位として評価するため，文中の語順が誤っている場合は正しい評価ができない．一般的に BLEU が 40 以上のときに高精度な翻訳が得られたと言われる．BLEU は式 (5.1) で求められる．BP (Brevity penalty: 短さペナルティ) は参照文に比べて出力文が短すぎる場合に与えられるペナルティであり，式 (5.2) で求められる．すなわち，出力文の長さ c が参照訳文の長さ r よりも短い場合に指数関数を用いた減衰率を求め，これを BP と定義する．出力文の長さ c が参照訳文の長さ r よりも長い場合は BP は 1 である．この BP を式 (5.1) 右辺の N-gram 適合率 p_n の重み付き和に乗じて BLEU を求める．

$$\mathrm{BLEU} = \mathrm{BP} \cdot exp(\sum_{n=i}^{N} w_n \log p_n) \tag{5.1}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{5.2}$$

CHRF^{++} CHRF^{++} (Character n-gram F-score^{++}) は文字 N-gram の適合率 $ngrP$ と再現率 $ngrR$ の調和平均を算出し，参照文との一致度を測る指標である [23, 28]．β は再現率に重みを与えるパラメータである．形態が豊富である日本語の翻訳の品質を測るため，CHRF^{++} の利用は適していると考えた．CHRF^{++} の計算式を式 (5.3) に示す．

$$\mathrm{CHRF}^{++} = (1+\beta^2)\frac{(ngrP \cdot ngrR)}{(\beta^2 \cdot ngrP + ngrR)} \tag{5.3}$$

5.2 実験の結果

各 NMT モデルをファインチューニングする際に使用した訓練データのサイズを表 5.2 に，それぞれの NMT モデルの自動評価の結果を表 5.3 に示す．コーパス拡張の際，反復的逆翻訳では 1 回の反復でコーパスのサイズが 2 倍になるはずだが，エラーなどで翻訳文が得られなかったことがあるため，コーパスのサイズに関しては単純な倍増よりも小さくなった．なお，表 5.3 に示した $M_{t\to s}$ は日本語→アイヌ語翻訳モデルの結果である．これらの翻訳モデルは反復的学習の過程で付随的に生成されたものである．本研究ではアイヌ語→日本語の機械翻訳を目的としており，日本語→アイヌ語の翻訳モデルは評価しないが，参考のため記載する．

表 5.2: 各モデルのファインチューニング時の訓練データのサイズ

反復的巡回翻訳		反復的逆翻訳 [9]	
Model	Size	Model	Size
$M^{base}_{s\to t}$	18,669	$M^{base}_{s\to t}$	18,669
$M^{C1}_{t\to s}$	37,245	$M^{B1}_{t\to s}$	18,669
$M^{C1}_{s\to t}$	73,814	$M^{B1}_{s\to t}$	36,877
$M^{C2}_{t\to s}$	164,310	$M^{B2}_{t\to s}$	72,696
$M^{C2}_{s\to t}$	233,605	$M^{B2}_{s\to t}$	127,183

表 5.3: 各モデルの自動評価結果

	ain → jpn			jpn → ain		
	Model	BLEU	CHRF^{++}	Model	BLEU	CHRF^{++}
ベースライン	$M^{base}_{s\to t}$	24.68	24.94	$M^{base}_{t\to s}$	24.04	**48.27**
反復的巡回翻訳	$M^{C1}_{s\to t}$	18.72	19.93	$M^{C1}_{t\to s}$	23.71	47.99
	$M^{C2}_{s\to t}$	15.73	18.24	$M^{C2}_{t\to s}$	23.42	47.92
反復的逆翻訳 [9]	$M^{B1}_{s\to t}$	**33.63**	**31.42**	$M^{B1}_{t\to s}$	n/a[1]	n/a[1]
	$M^{B2}_{s\to t}$	25.34	26.41	$M^{B2}_{t\to s}$	**24.19**	47.64

[1] $M^{base}_{t\to s}$ と同じモデルのため結果掲載省略

5.3 考察

5.3.1 自動評価の考察

反復的逆翻訳をアイヌ語→日本語翻訳モデルに適用したところ，n=1 のときに最も評価値が改善し，初期パラレルコーパスのみを用いてファインチューニングを行ったベースラインのモデルと比較すると，BLEU スコアが+8.95，CHRF^{++} スコアが+6.48 向上した．n=2 のときには n=1 のときと比べて両評価値が低下した．

一方，反復的巡回翻訳を適用したところ，反復回数を増やすたびに評価値が低下し，ベースラインのモデルと比較して精度向上は確認できなかった．

反復的逆翻訳においては n=1 のときに改善が確認されたが，n=2 のとき，ならびに反復的巡回翻訳適用時には反復を重ねるたびに評価値が低下したことを鑑みると，アイヌ語から日本語への翻訳モデルを反復学習によって獲得する際には，単に反復回数を増やすだけでは翻訳の性能は向上せず，拡張データの品質を高めるなどさらなる工夫が必要である．

5.3.2 翻訳例の分析

人手評価として，翻訳例の一部を人手で確認し，誤りの要因を分析する．自動評価では反復的逆翻訳の n=1 のときに評価値が改善したが，実際の出力文を確認しても，n=1 と比べて n=2 のモデルの出力の方が崩れた訳出が多かった．また反復的巡回翻訳で得られたモデルの出力文は反復的逆翻訳で得られたモデルと比べて，表現面で見劣りする場合が多く，誤訳も多かった．

以下，翻訳の例を 7 つ挙げる．例 1，例 2，例 3 を表 5.4 に，例 4，例 5，例 6，例 7 を表 5.5 に掲載する．

表 5.4: 各モデルによる翻訳の例 (1)

例 1:

原言語文	ku= kema pop us
目標言語文	私の足にまめができた.
$M_{s\to t}^{base}$	私の足に生えている.
$M_{s\to t}^{C1}$	足が落ちてしまいました.
$M_{s\to t}^{C2}$	私はそれを作りました.
$M_{s\to t}^{B1}$	私の足が流行りそうだ.
$M_{s\to t}^{B2}$	私は足が速い.

例 2:

原言語文	tanpe aynu or un ku= ye eaykap
目標言語文	これをアイヌ語で言えません.
$M_{s\to t}^{base}$	これは人間のところで言うことはできない.
$M_{s\to t}^{C1}$	私はこれを人間に言うことができません.
$M_{s\to t}^{C2}$	これは人間には言えません.
$M_{s\to t}^{B1}$	これは人間のところで言うことができません.
$M_{s\to t}^{B2}$	これはアイヌのところに言うことができません.

例 3:

原言語文	e= i= tuye yakka taa koraci kamuy an= ne kus siknu =an na
目標言語文	おまえが私を切ってもこのように私は神なので生きられるのだ.
$M_{s\to t}^{base}$	あなたは私を切ってもこのように私は神なので生きられるのだよ.
$M_{s\to t}^{C1}$	あなたが私を切っても，私はこのように神なので生きます.
$M_{s\to t}^{C2}$	あなたが私を切っても，私はこのように神なので生きることができます.
$M_{s\to t}^{B1}$	あなたは私を切ってもこのように私は神なので生きていますよ.
$M_{s\to t}^{B2}$	あなたが私を切ってもこのように私は神なので生きられますよ.

表 5.5: 各モデルによる翻訳の例 (2)

例 4:

原言語文	onne kusu ne hi ta ne sioroomare okkaypo utar yupyupkeno kaspaotte
目標言語文	死ぬという時にその同居させた若い男達にきつく言い聞かせました.
$M_{s\to t}^{base}$	年を取った時にその集まった若者たちはきつく言い聞かせました.
$M_{s\to t}^{C1}$	年を取ったとき，例の若い男性たちは言った.
$M_{s\to t}^{C2}$	私が年をとったとき，私は若い男たちに言い聞かせました.
$M_{s\to t}^{B1}$	死にそうになった時にその家に入っていた若い男性たちは言い負かされひどく懲らしめられました.
$M_{s\to t}^{B2}$	年を取ろうとした時にその座っている若者たちは言い聞かせました.

例 5:

原言語文	ora ne a p ka nitne inaw poka e= koroski yak pirka na
目標言語文	そんなものにでも魔神に捧げるイナウだけは立てるといいですよ.
$M_{s\to t}^{base}$	そしてその物も粗末な木幣だけでも供えてあげてください.
$M_{s\to t}^{C1}$	しかし，あなたはそれを木で削った木で与えます.
$M_{s\to t}^{C2}$	それなのに粗末な木幣だけでも持ってきてください.
$M_{s\to t}^{B1}$	そのものも粗末な木幣だけでも供えてください.
$M_{s\to t}^{B2}$	それなのに粗末な木幣だけでももらってあげたらいいよ.

例 6:

原言語文	penampe anakne sonno yayasis kor wen ray toy ray ki wa isam
目標言語文	ペナンペは心から悔みながら悪い死つまらぬ死を遂げてしまった.
$M_{s\to t}^{base}$	ペナンペは本当にお腹が空いて悪い死に方をして死んでしまった.
$M_{s\to t}^{C1}$	ペナンペはひどくかしこまってひどく死んでしまった.
$M_{s\to t}^{C2}$	ペナンペは本当に後悔してひどく死にました.
$M_{s\to t}^{B1}$	ペナンペはひどく悔しくてひどく死んでしまいました.
$M_{s\to t}^{B2}$	ペナンペはひどく悔しく思いながらひどく死んで死に絶えました.

例 7:

原言語文	tu onkami ne re onkami onkami kane yayrayke
目標言語文	何度も拝礼を繰り返して感謝しました.
$M_{s\to t}^{base}$	二つの礼拝，三つの礼拝をしながら感謝します.
$M_{s\to t}^{C1}$	何度も拝礼をして感謝しました.
$M_{s\to t}^{C2}$	2 度も 3 度も拝礼をして感謝しました.
$M_{s\to t}^{B1}$	ふたつの礼拝，三つの礼拝などで感謝しました.
$M_{s\to t}^{B2}$	何度も拝礼をして感謝しました.

例 1 では，pop us (*jpn.* まめができる) という表現の訳出ができていなかった．反復的巡回翻訳では「私の」という所属表現が抜け落ちてしまっており，また n=2 の時の NMT モデル $M^{C2}_{s\to t}$ の出力では kema (*jpn.* 足) という一般的な語も訳出できていない．

表 5.6 に，表 5.4 例 1 の翻訳ペアに対し，データ拡張によって得られた翻訳ペアの変遷を示す．TP-ID はデータ拡張によって生成された翻訳ペアの ID である．ID は「T^n_X–N」の形式で，T^n_X は図 4.2 に示した拡張パラレルコーパスの記号 ($X = \{a, b\}$) を，N は識別番号を表す．

目標言語文の巡回翻訳による拡張後の日本語 T^1_a の時点で，pop us の対訳例として「立ち上がっていました」と誤訳の翻訳ペアが生成されており，pop us が日本語で「まめができる」を意味することを学習するための翻訳ペアが生成できていなかった．. 巡回翻訳によるコーパス拡張の結果，適切な翻訳ペアが追加できていれば，次の手順である逆方向翻訳モデル $M^{C1}_{t\to s}$ の学習を補助することができた可能性がある．

例 2 はアイヌ語文を語義通り直訳すると「これをアイヌのところでは私は言うことが出来ない」という文になるが，この表現で「アイヌ語では表現できない」という意味を表す．aynu はアイヌ語で「人間」と「アイヌ民族」という 2 つの意味があり，訓練データ中で「人間」と和訳される場合が多かったのか，aynu の「アイヌ民族」の意味が正しく認識されなかった可能性がある．or un は「〜へ，〜のところで」という意味で動作の向かっていく方向を示す表現だが [22]，aynu or un で「アイヌ語で」という意味を表す慣用的な表現である．この or un もその他の用例の影響から，慣用表現の訳出ができず，結果として「人間のところで」「アイヌのところに」という訳出になったケースが多かった可能性がある．語義通りの直訳から鑑みると，ベースラインモデル $M^{base}_{s\to t}$ と反復的逆翻訳の n=1 のときのモデル $M^{B1}_{s\to t}$ の出力は「人間のところで言うことができない」と訳出しており許容範囲内だが，他のモデルではその表現すら出力できておらず，反復学習を重ねることで語義通りの訳出もできなくなっている．

また，主語，目的語の関係が正しく認識されていないケースが多く，課題として残されている．例 3 は，e=i=tuye yakka (*jpn.* あなたが-我を-切る-するとしても) のように，二人称主格接辞 (e=) ，四人称目的格接辞 (i=) が tuye (*jpn.* 切る) という二項動詞に接合されており，このアイヌ語文の主語，目的語の訳出は接続されている接辞に基づいていずれのモデルも正しく出力できていた．一方で，例 4 では，kaspaotte (*jpn.* 諭す) という二項動詞が用いられている．二項動詞はその名の通り，主語と直接目的語の 2 項を取るため，前後の文脈が提示されていない環境における解釈としては，三人称を主語に，ne sioroomare okkaypo utar (*jpn.* その同居させた男たち) を目的語にとるか，もしくは ne sioroomare okkaypo utar を主語に，三人称を目的語にとるかの 2 択となる．この例では前者が正解であるが，本実験における各モデルの出力では後者で訳出されたものが多かった．反復的巡回翻訳の n=2 のときのモデル $M^{C2}_{s\to t}$ の出力では，入力文中に存在しない「私

表 5.6: 翻訳ペア拡張の変遷: 表 5.4 の例 1 の翻訳ペアを例に

入力文	翻訳モデル	TP-ID	アイヌ語文	日本語文
D^p	-	-	ku= kema pop us	私の足にまめができた.
n=1: 目標言語-巡回翻訳				
D^p の日本語文	$M_{t\to a}$ & $M_{a\to t}$	T_a^1	ku= kema pop us	**私は立ち上がっていました.**
n=1: 原言語-逆翻訳				
D^p の日本語文	$M_{t\to s}^{C1}$	T_b^1–1	**an= cikiri cikoykip kor**	私の足にまめができた.
T_a^1 の日本語文	$M_{t\to s}^{C1}$	T_b^1–2	**hopunpa =an**	私は立ち上がっていました.
n=1: 目標言語-順翻訳				
D^p のアイヌ語文	$M_{s\to t}^{C1}$	T_c^1–1	ku= kema pop us	**足が落ちてしまいました.**
T_b^1–1 のアイヌ語文	$M_{s\to t}^{C1}$	T_c^1–2	an= cikiri cikoykip kor	**私の足は獲物を持ってきました.**
T_b^1–2 のアイヌ語文	$M_{s\to t}^{C1}$	T_c^1–3	hopunpa =an	**私は立ち上がった.**
n=2: 目標言語-巡回翻訳				
T_c^1–1 の日本語文	$M_{t\to a}$ & $M_{a\to t}$	T_a^2–1	cikirihi hacir	**私の足が落ちました.**
T_c^1–2 の日本語文	$M_{t\to a}$ & $M_{a\to t}$	T_a^2–2	cikirihi cikoykip kor wa ek	**私の足は彼らの獲物をもたらしました.**
n=2: 原言語-逆翻訳				
D^p の日本語文	$M_{t\to s}^{C2}$	T_b^2–1	**cikirihi a= yaykotuwaskarap**	私の足にまめができた.
T_a^1 の日本語文	$M_{t\to s}^{C2}$	T_b^2–2	**hopunpa =an**	私は立ち上がっていました.
T_c^1–1 の日本語文	$M_{t\to s}^{C2}$	T_b^2–3	**cikiri a= turseka**	足が落ちてしまいました.
T_c^1–2 の日本語文	$M_{t\to s}^{C2}$	T_b^2–4	**cikirihi cikoykip kor wa ek**	私の足は獲物を持ってきました.
T_c^1–3 の日本語文	$M_{t\to s}^{C2}$	T_b^2–5	**hopunpa =an**	私は立ち上がっていました.
T_a^2–1 の日本語文	$M_{t\to s}^{C2}$	T_b^2–6	**cikirihi hacir**	私の足が落ちました.
T_a^2–2 の日本語文	$M_{t\to s}^{C2}$	T_b^2–7	**cikirihi cikoykip kor wa ek**	私の足は彼らの獲物をもたらしました.

は」と言う主語を追加しており，文義が大きく変わってしまった．これに加えて，いずれのモデルも sioroomare (< si-oro-oma-re 自分-のところに-入れる-使役形化接尾辞，*jpn.* 同居させる) という語の訳出ができていなかった．sioroomare は初期コーパス 23,337 文のうち 2 文 3 回のみ出現した低頻度語である．反復的逆翻訳の n=1 のときのモデル $M^{B1}_{s\to t}$ では「家に入っていた」と正しくはないが近い表現で訳出していたことを考えると，反復を重ねたことでパラレルコーパス全体においてさらに相対的な出現頻度が下がり，正しい学習ができなくなったと推測される．反復的巡回翻訳を適用した 2 つのモデルでは sioroomare の訳出ができておらず，他のテスト文でも同様に低頻度語を訳し落としてしまうことが多かった．

文法以外にはアイヌ文化関連語の認識が足りていない点が見られた．例 5 は，inaw (*jpn.* イナウ，木幣) と呼ばれる，アイヌ民族の祭具・供物の単語が含まれる文である．nitne inaw で「魔神 (*ain.* nitne kamuy) に捧げるイナウ，粗末なイナウ」と言う意味を持つが，反復的巡回翻訳の n=1 のときのモデル $M^{C1}_{s\to t}$ の出力では,「木で削った木」といった本来の意味と全く異なる表現に翻訳されていた．巡回翻訳時に用いた日英・英日翻訳システムはアイヌ文化関連語に特化しているわけではないため，巡回翻訳により拡張された日本語文によって誤った翻訳ペアが生成され，これが翻訳モデルの学習に悪影響を与えた可能性がある．また，既存手法モデルの $M^{B2}_{s\to t}$ を除いて，yak pirka (*jpn.* 〜すると良い) という頻出表現の訳出が適切にできていなかった．特に反復的巡回翻訳の n=1 のときのモデル $M^{B1}_{s\to t}$ は大部分が誤訳となっており，yak (*jpn.* したならば) すら訳出できていなかった．頻出表現であるにも関わらず翻訳された表現にばらつきがあることを考慮すると，この表現でも inaw の訳出のときと同様に，拡張された誤った翻訳ペアが翻訳モデルの学習に悪影響を与えた可能性がある．生成された全ての翻訳ペアをパラレルコーパスに追加するのではなく，品質の高い翻訳ペアのみ追加するなどして，データ拡張の量や質を調整する必要がある．

例 6 は，“wen ray toy ray ki wa isam” (*jpn.* 悪い死つまらぬ死を遂げてしまった) のように，wen (*jpn.* 悪い) と toy (*jpn.* ひどい) という二つの形容詞が ray (*jpn.* 死) という名詞を修飾しており，Adj_1 NP Adj_2 NP[2]という対句表現 (Adj_1, Adj_2 はそれぞれ別の語) をなしている．また，この対句表現に加えて，ki (*jpn.* する) という動詞を使い,「悪い死，ひどい死をしてしまった」と意味をなす迂言的な表現の訳出例である．これらはアイヌ語の雅語 (*ain.* a=tomte itak) で頻出する表現である [22]．この対句や迂言的な表現の訳出だけに注目すると，概ね許容範囲内の出力とみなせる．雅語表現は普段の言葉と異なる特別な修辞を用いる点で，出力文もその修辞が反映されていると尚良いと思われる．その点，反復的逆翻訳の n=2 のときのモデル $M^{B2}_{s\to t}$ の出力のような重言表現は差し支えないと思われるが，例えば反復的巡回翻訳の n=2 のときのモデル $M^{C2}_{s\to t}$ の出力文はそういった特有の修辞が表現されておらず，印象として淡白である．文義自体は伝達できるとしても，アイヌ口承文芸における修辞法を日本語で表現する際に NMT モデルを用い

[2]Adj = 形容詞，NP = 名詞句

る場合は，そのモデルがニュアンスを適切に訳出できることが求められる．

最後の例 7 は，“tu onkami ... re onkami” (*jpn.* 何度も拝礼をして) という対句表現の例である．アイヌの口承文芸では，tu NP re NP (*jpn.* 2 つの NP，3 つの NP → たくさんの，何度も) は例 6 の表現同様に頻出する表現である．反復的巡回翻訳の n=1 のときのモデル $M^{C1}_{s\to t}$ と反復的逆翻訳の n=2 のときのモデル $M^{B2}_{s\to t}$ では適切に訳出ができていた．反復的逆翻訳の n=1 のときのモデル $M^{B1}_{s\to t}$ では「ふたつの... 三つの...」と訳出していたことを考えると，反復的逆翻訳においては反復を重ねることで，迂言的な表現や対句表現の翻訳ペアの拡充がある程度達成できていたと考えられる．一方，反復的巡回翻訳では反復を重ねたことで，対句表現を正しく翻訳する能力が失われてしまった．語義通りの訳出はできているものの，迂言的な表現や対句表現の翻訳にはまだ改善の余地が残されていることが示唆された．

5.4 問題点のまとめ

自動評価では，本研究で提案した反復的巡回翻訳は既存の反復的逆翻訳よりも BLEU や chrF^{++} が低かった．また，実際の翻訳例を確認したところ，いくつかの問題点が明らかになった．本節では反復的巡回翻訳の問題点を整理する．

学習データの質的・量的な問題 本稿で利用したアイヌ語の文献数自体が少なく，また初期パラレルコーパスの日本語文も表現上わかりにくいことがあり，改善の余地があった可能性がある．こういった質的・量的な制約条件のもとで目標言語文の巡回翻訳によるデータ拡張を行ったことで，拡張された文の表現が限定的だったり，好ましくない表現が生成された可能性がある．また文献量に加え，例 4 で示したような低頻度語の翻訳にも課題が残る．アイヌ語，日本語の単語の種類は非常に多いため，限られたパラレルコーパスでは適切な翻訳モデルが学習できなかった可能性がある．

データのドメイン・方言ごとの分離 本研究ではデータ数を確保するため，方言やドメイン (口承文芸・会話体) を区別せずに一つのパラレルコーパスを作成したが，異なるドメイン・方言の翻訳モデルを単一のモデルとして学習したことにより，結果としてそれぞれのドメイン・方言における翻訳の特徴を学習できなかった可能性がある．

主語・目的語の逆転 例 4 で示した通り，動詞の取りうる項数が動詞ごとに定まっているのにもかかわらず，それが認識されておらず，主語と目的語の順番が逆転した出力文が存在した．アイヌ語の場合，三人称は無表示であり，かつ日本語文側でも文脈を踏まえない限りその主語と目的語の関係がわからない翻訳ペアが存在したため，主語と目的語の順序に関する知識が学習できず，誤った文が生成される要因となった可能性がある．

慣用表現・雅語表現の訳出 例6で示した通り，アイヌ語では口承文芸の対句表現や婉曲表現が用いられることがある．こういったアイヌ語特有の表現の対訳として，初期パラレルコーパスの日本語文では逐語訳されている文と意訳されている文があり，表現に揺らぎがあった．これにより散文表現の翻訳に関して適切な学習ができなかった可能性がある．一方で提案手法による雅語表現の拡充によって翻訳の品質が向上したケースも確認された．

目標言語-補助言語間の翻訳モデルの精度 本研究では目標言語 (日本語) 文拡張の際の巡回翻訳にGoogle翻訳を利用したが，Google翻訳はアイヌ文化関連語を含む文章の翻訳に特化しているわけではない．そのため補助言語を介した巡回翻訳を経て目標言語文の意味が大きく変化した可能性がある．

第6章　おわりに

6.1　まとめ

本研究では，低資源言語である北海道アイヌ語から日本語への機械翻訳を対象に，反復的逆翻訳によって翻訳の性能を高めることができるかを検証した．反復的逆翻訳では，初期の少量のパラレルコーパスから逆翻訳モデル (日本語→アイヌ語) を学習し，これを用いてパラレルコーパスの目標言語の文を原言語の文に翻訳することで新たな翻訳ペアを得て，これをパラレルコーパスに追加した．次に，拡張したパラレルコーパスを用いて翻訳モデル (アイヌ語→日本語) を学習し，これを用いてパラレルコーパスの原言語の文を目標言語の文に翻訳することで新たな翻訳ペアを得て，パラレルコーパスに追加した．上記の一連のコーパス拡張とNMT モデル学習の操作を反復することで，パラレルコーパスを漸進的に拡張し，かつ NMT モデルの品質を漸進的に向上させた．さらに，上記の反復的逆翻訳を改良した「反復的巡回・逆・順翻訳」を提案した．逆翻訳によるコーパス拡張の前に，巡回翻訳によるコーパス拡張の処理を追加した．比較的精度の高い既存の日英・英日機械翻訳システムを用いて，パラレルコーパスにおける目標言語の文 (日本語文) を英語文に翻訳し，これを日本語文に再度翻訳することで新たな翻訳ペアを得た．

評価実験では，反復的逆翻訳によって翻訳の精度が向上することを確認した．初期のパラレルコーパスのみから学習されたベースラインモデルに比べて，反復回数が 1 回のときには BLEU スコアが 8.95 ポイント向上した．ただし，反復回数を 2 回に増やすと，BLEU スコアはベースラインモデルに比べると微増したが，反復回数 1 回のときよりも低下した．一方，本研究で提案した反復的巡回・逆・順翻訳を適用したとき，ベースラインモデルに比べて，反復回数が 1 回の時は BLEU スコアが 5.96 ポイント低下し，反復回数が 2 回の時は BLEU スコアが 8.95 ポイント低下した．反復回数を重ねるたびに評価値が低下し，改善はみられなかった．さらに，いくつかの翻訳例に対して詳細なエラー分析を行い，提案手法を改善するためのいくつかの方策を明らかにした．

6.2　今後の課題

本研究の今後の課題を以下に述べる．

文献数の質的・量的問題の改善 今回の実験で使用していないアイヌ語の文献，特に電算化されていない文献をさらに追加することで，初期のパラレルコーパスの量が少ないという問題を改善できる可能性がある．また，質的な問題の改善として，本研究で利用したテキストの表記は原文献のままとしたが，一部の文献では句点が記載されていないことが多く，結果として一つの文が非常に長くなったケースがあった．これにより，NMT モデルを適切に学習できなかった可能性があるため，パラレルコーパス内の文の区切りを修正することが必要である．また，上記の質的問題とは別に，原文献の日本語訳が逐語訳になっているものがあり，一般的な日本語の語順ではないケースも存在した．パラレルコーパスの日本語訳の文を自然なものに書き換えることで翻訳の精度が向上するか検証することも必要であろう．

口承文芸と会話文，方言間の分離 ブガエワが言及しているように [3]，アイヌ語の文献の多くは口承文芸が多く，会話文の文献数は限られる．口承文芸と会話文で用いられる人称接辞や語彙，表現は異なるため，Igarashi と Miyagawa が実施しているように [11]，口承文芸と会話体の文献を分けることで，実際の会話で用いられるより適切な表現を得ることができると思われる．また，方言差に関して，第 3 章で言及した通り，本研究で整備したパラレルコーパスはデータ数を確保するために方言間の差異を考慮していない．しかし，例えば沙流方言話者の会話文中に別地域の語彙が含まれることは，実際の言語利用を考慮するとあまり好ましいとは言えない．そのため，先行研究 [11] のように方言ごとにコーパスを分ける必要もある．文献数を確保することにより，方言ごとに分けた場合でもある程度のデータ数が確保できたとき，本研究の提案手法によって翻訳精度が同様に向上するか検証することが必要であろう．

逆翻訳後のアイヌ語の非文検査: 項数のチェック機能 逆翻訳を用いてコーパス拡張を行う際，日本語からアイヌ語への翻訳によってアイヌ語文の非文が生成される可能性がある．アイヌ語文の非文検査の一つとして，アイヌ語動詞の項数チェックを利用する方法が考えられるだろう．アイヌ語の動詞はその動詞が持ちうる項が定められているため [22]，例えば動詞と項数のデータが記載された辞書を参照し，翻訳後のアイヌ語文の項数を検査する機構を追加することで，アイヌ語文の品質を担保できるのではないかと考えている．本研究においてはアイヌ語動詞の項数を含む辞書を作成する時間がなかったが，アイヌ語機械翻訳の精度向上の一案として今後の課題としたい．

補助言語資料の拡張 本研究においては目標言語 (日本語) から補助言語 (英語) への翻訳に Google 翻訳を用いたが，Google 翻訳はアイヌ文化関連語を含む文章の翻訳に特化しているわけではない．そこで，アイヌ文化関連語を含む日英文献を訓練データとし，低言語資源の言語の翻訳むけに事前学習されたモ

デルである NLLB (No Language Left Behind) [29] をベースモデルとして用いて日英方向・英日方向の NMT モデルをファインチューニングした．しかし，ファインチューニングに使用した文献が少なかったため，十分に高い翻訳精度が得られなかった．今後，アイヌ語・日本語パラレルコーパスのみでなく，補助言語資料として英語訳の拡張を実施し，目標言語・補助言語間の翻訳モデルを学習したとき，原言語から目標言語方向への翻訳精度が向上するかを検証することは重要な課題である．

参考文献

[1] Tucker Berckmann and Berkan Hiziroglu. Low-resource translation as language modeling. *Proceedings of the 5th Conference on Machine Translation (WMT)*, pp. 1079–1083, 2020.

[2] 平取町立二風谷アイヌ文化博物館. アイヌ語・アイヌ口承文芸. http://www.town.biratori.hokkaido.jp/biratori/nibutani/culture/language/story/. 2023 年 10 月 31 日閲覧.

[3] ブガエワアンナ. 北海道南部のアイヌ語. 早稲田大学高等研究所紀要, Vol. 6, pp. 33–76, 2014.

[4] 国立大学法人 千葉大学 大学院人文社会科学研究科. アイヌ語鵡川方言 日本語－アイヌ語辞典. https://www.gshpa.chiba-u.jp/cas/Ainu-archives/index.html. 2024 年 7 月 8 日閲覧.

[5] 知里真志保. アイヌ民譚集. 岩波書店, 1981.

[6] 越前谷博, 荒木健治, 桃内佳雄, 栃内香次. 局所着目方式によるアイヌ語ー日本語名詞対訳語の自動抽出について. 情報処理学会 研究報告, pp. 161–166, 2003.

[7] 越前谷博, 荒木健治, 桃内佳雄. アイヌ語ー日本語対訳コーパスを対象とした局所着目型学習による対訳語の自動抽出. 北海学園大学工学部研究報告, pp. 41–63, 2005.

[8] Isaac Feldman and Rolando Coto-Solano. Neural machine translation models with back-translation for the extremely low-resource indigenous language bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3965–3976, 2020.

[9] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18–24, 2018.

[10] 社団法人北海道ウタリ協会. アコロ イタク：アイヌ語テキスト 1. クルーズ, 1994.

[11] Ryo Igarashi and So Miyagawa. Enhancing neural machine translation for ainu-japanese: A comprehensive study on the impact of domain and dialect integration. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 413–422, 2024.

[12] 片山龍峰. 1991 年 11 月 10 日 千歳市蘭越生活館におけるアイヌ語会話ビデオ. 未公開資料, 1991.

[13] 萱野茂. 萱野茂のアイヌ語辞典. 三省堂, 2002.

[14] Ahrii Kim and Jinhyeon Kim. Vacillating human correlation of sacrebleu in unprotected languages. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 1–15, 2022.

[15] 国立アイヌ民族博物館, 公益財団法人アイヌ民族文化財団. 国立アイヌ民族博物館アイヌ語アーカイブ. `https://ainugo.nam.go.jp/`. 2024 年 7 月 8 日閲覧.

[16] 国立国語研究所. トピック別アイヌ語会話辞典. `https://ainu.ninjal.ac.jp/topic/dictionary/jp/`. 2024 年 7 月 8 日閲覧.

[17] 久高優也. 琉日機械翻訳のための対訳コーパスの自動拡張について. Master's thesis, 北陸先端科学技術大学院大学, 2019.

[18] So Miyagawa. Machine translation for highly low-resource language: A case study of ainu, a critically endangered indigenous language in northern japan. *Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL*, pp. 120–124, 2023.

[19] 桃内佳雄, 大友雄介, 越前谷博. アイヌ語・日本語機械翻訳のための基礎的考察. 言語処理学会 第 8 回年次大会 発表論文集, pp. 25–28, 2002.

[20] 中川裕, ブガエワアンナ, 小林美紀, 吉川佳見. アイヌ語口承文芸コーパス―音声・グロス付き―, 2016–2024. `https://ainu.ninjal.ac.jp/folklore/corpus/jp/`. 2024 年 7 月 8 日閲覧.

[21] 中川裕 (補訂). 知里 幸惠 アイヌ神謡集. 岩波書店, 2023.

[22] 中川裕. アイヌ語広文典. 白水社, 2024.

[23] Maja Popović. chrF^{++}: words helping character n-grams. In *Proceedings of the Conference on Machine Translation*, Vol. 2, pp. 612–618, 2017.

[24] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, Vol. 1, pp. 186–191, 2018.

[25] Michael Przystupa and Muhammad Abdul-Mageed. Neural machine translation of low-resource and similar languages with backtranslation. *Proceedings of the Fourth Conference on Machine Translation (WMT)*, pp. 224–235, 2019.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[27] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *CoRR*, 2016.

[28] 嶋中宏希, 梶原智之, 小町守. 事前学習された文の分散表現を用いた機械翻訳の自動評価. 自然言語処理, Vol. 26, No. 3, pp. 613–634, 2019.

[29] NLLB Team, Marta R. Costa-jussá, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *Computation and Language*, 2022.

[30] Bui Tuan Thanh, 秋葉友良, 塚田元. 双方向翻訳モデルと反復的逆翻訳を用いた低資源言語に対するニューラル機械翻訳の性能向上. 言語処理学会 第 28 回年次大会 発表論文集, pp. 1756–1760, 2022.

[31] 東京外国語大学アジア・アフリカ言語文化研究所. 情報資源利用研究センター・プロジェクト「アイヌ語音声資料の文字化テキスト対応づけと公開」. `https://ainugo.aa-ken.jp/main.php?id=1`. 2024 年 7 月 8 日閲覧.

[32] Edward (ed.) Vajda. The ainuic language family. *The Languages and Linguistics of Northern Asia*, Vol. 1, pp. 541–632, 2024.

[33] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 856–863, 2007.

[34] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, 2021.

[35] Leng Yichong, Tan Xu, Qin Tao, Li Xiang-Yang, and Tie-Yan Liu. Unsupervised pivot translation for distant languages. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 175–183, 2019.

[36] 全国大学生活協同組合連合会. わが大学の先生と語る「『金カム』監修者と語るアイヌ」, 2022. 2022 年 10 月 4 日閲覧.