逆翻訳を用いたアイヌ語・日本語機械翻訳の改善における研究

菅原 葵[1], Karol Nowakowski[1], Michal Ptaszynski[2], Nick Overacker[2]
[1]東北公益文科大学 公益学部 公益学科 メディア・情報コース
[2]北見工業大学 工学部 地域未来デザイン工学科 情報デザイン・コミュニケーション工学コース
aoimizutama0@gmail.com karol@koeki-u.ac.jp
michal@mail.kitami-it.ac.jp m3235380013@std.kitami-it.ac.jp

概要

本研究では，逆翻訳を用いて少資源言語であるアイヌ語と日本語の機械翻訳の精度向上を目指した．元の対訳コーパスのみを利用して，反復的逆翻訳手法とランダムサンプリングによる疑似対訳文生成手法を用い，さらに，外部の日本語の単言語データを活用した．結果として，外部の日本語の単言語データを利用するよりも元の対訳コーパスのみを利用して逆翻訳を行った方が良い結果となった．

1 はじめに

UNESCO [1]によると，世界で話されている言語のうちの約 40%が消滅の危機にあるとされている．アイヌ語もその 1 つであり，極めて深刻な消滅の危機にあるとされている [2]． また機械翻訳をはじめとする自然言語処理では，アイヌ語のような少資源言語の研究は，データが多い言語と比べ遅れていて精度が低い現状にある．

少資源言語の機械翻訳の精度を向上させる手法として，比較的入手が容易な単言語コーパスを利用し，逆翻訳 [3]を用いてデータを拡張する手法は効果的な手法として知られている．また Thanh ら [4]は，元の対訳コーパス以外の追加の学習データを使わずに逆翻訳を行う手法を提案した．また近年では，事前学習済みのトランスフォーマーモデルの利用が，アイヌ語を含む少資源言語の翻訳精度の向上につながることがわかった [5].

本研究では多言語事前学習済みモデルの機械翻訳モデル（NLLB-200 [6]）および Thanh らの元の対訳コーパスを用いた逆翻訳手法を利用し，さらに外部の日本語の単言語データも利用することで，アイヌ語と日本語の機械翻訳精度をより向上させる可能性を検討する．

2 関連研究

2.1 アイヌ語と日本語の機械翻訳

Miyagawa [7]は Marian MT [8]を利用してアイヌ語と日本語のニューラル機械翻訳モデルを構築した．田中ら [5]はアイヌ語と日本語のデータを用いて Bilingual BERT モデルを事前学習し，その出力である単語埋め込みをトランスフォーマーに基づく機械翻訳モデルに適用する手法を提案した．

2.2 逆翻訳

Sennrich ら [3]は単言語データを学習済みのモデルで翻訳して疑似の対訳コーパスを作成し，ニューラル機械翻訳モデル学習において単言語データを活用する「逆翻訳」の手法を提案した．Hoang ら [9]と森田ら [10]は双方向機械翻訳モデルの開発において，原言語側と目的言語側の単言語データを用いて逆翻訳を繰り返すことで，反復的にモデルを改善する手法を提案した．今村ら [11]は疑似原文を生成する際，ランダムサンプリングによって複数の文を生成し疑似対訳文の多様性を増大させる手法を提案した．Thanh ら [4]は双方向機械翻訳モデルとサンプリングを用いた反復的逆翻訳を組み合わせ，元の対訳コーパス以外の追加の学習データを使わずに翻訳性能を向上させることに成功した．

3 提案手法

3.1 NLLB-200 モデル

本研究では NLLB-200 というマルチリンガル事前学習済み機械翻訳モデルを利用してアイヌ語と日本語の機械翻訳モデルを開発した．先行研究 [4] [7]では少資源機械翻訳における双方向学習の有効性が実証されたため，本研究においてもアイヌ語-日本語

および日本語-アイヌ語の双方向で同時に学習を行った．また，NLLB の「言語タグ」にはアイヌ語を表すタグは含まれていなかったため，アイヌ語専用の言語タグを新たに追加した．

3.2 元の対訳コーパスの反復的逆翻訳

アイヌ語は単言語データの獲得が難しいため，本研究では Thanh ら [4]と同様に，元の訓練データおよびサンプリングを用いた反復的逆翻訳を利用した．このアプローチによるアイヌ語と日本語の翻訳モデル構築の手順はアルゴリズム 1 と図 1 に示す．

アルゴリズム 1　元の対訳コーパスの反復的逆翻訳とランダムサンプリングを用いたアイヌ語と日本語の翻訳モデルを構築する手順

(ア) 初期設定：

1: **n** ← 反復する回数（逆翻訳する回数＋1）
2: **i** ← 1
3: **k** ← 生成する疑似対訳文の個数
4: **全体の学習ステップ数** ← 500,000
5: **一度の学習ステップ数** ← 全体の学習ステップ数 ÷ n
6: **元の訓練データ** ← 元の対訳コーパスの 80%を無作為抽出した結果
7: k が 2 以上であれば，元の訓練データをオーバーサンプリングする：
8: 　**元の訓練データ** ← 元の訓練データ × k
9: **アイヌ語-日本語の疑似対訳データ** ← ∅
10: **日本語-アイヌ語の疑似対訳データ** ← ∅
11: **翻訳モデル** ← ‘nllb-200-distilled-600M’

(イ) 以下の操作を，i が n になるまで繰り返す：

1: i が 2 以上であれば：
2: 　**アイヌ語-日本語の疑似対訳データ** ← 翻訳モデルおよびランダムサンプリングを用いて，元の訓練データの日本語側をアイヌ語へ翻訳した結果
3: 　**日本語-アイヌ語の疑似対訳データ** ← 翻訳モデルおよびランダムサンプリングを用いて，元の訓練データのアイヌ語側を日本語へ翻訳した結果
4: **アイヌ語-日本語の訓練データ** ← 元の訓練データ ⊕ アイヌ語-日本語の疑似対訳データ
5: **日本語-アイヌ語の訓練データ** ← 元の訓練データ ⊕ 日本語-アイヌ語の疑似対訳データ
6: **今回の訓練データ** ← アイヌ語-日本語の訓練データ ⊕ 日本語-アイヌ語の訓練データ
7: **翻訳モデル** ← 今回の訓練データを用いて，翻訳モデルのパラメータ更新を，一度の学習ステップ数だけ行い，結果として得られたチェックポイントの中から，検証データでの BLEU スコアが最も高いもの
8: **i** ← i＋1

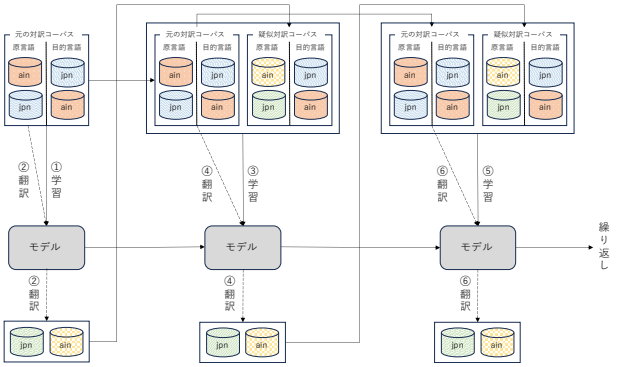

図 1　反復的逆翻訳とサンプリングを利用したアイヌ語と日本語の機械翻訳モデルの学習の流れ

3.3 外部の日本語単言語データの利用

アイヌ語はデータ資源が少ないが日本語は多くのデータが存在する．また，Sennrich ら [3]の研究において，逆翻訳するデータとして元の訓練データよりも新たな単言語データを用いた方が大幅な精度向上につながることが報告された．そこで元の対訳コーパスに含まれている日本語文以外の単言語データを利用する実験も行った．具体的には図 2 のように，アイヌ語-日本語の疑似対訳データを生成する際に，元の訓練データの日本語側の代わりに外部のコーパスから得られた日本語データを用いた．なお，外部のコーパスを利用する場合は，サンプリングを実行せずとも，元のデータになかった新たな文を追加すること自体が学習データの多様性の増大につながる．また，疑似原文を生成する際に，最大確率の出力を選択する「1 ベスト生成」を行う場合は，サンプリングによる生成と比べより高い翻訳品質が期待できる [11]．そのため，外部の単言語データを逆翻訳する際はランダムサンプリングを使用しなかった．

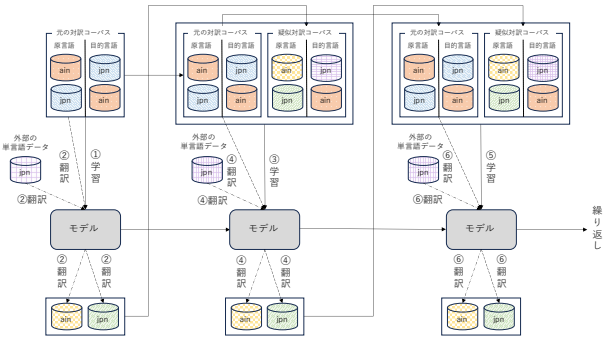

図 2　外部の日本語の単言語データを利用したアイヌ語と日本語の機械翻訳モデルの学習の流れ

4 実験

4.1 データ

本研究においてアイヌ語と日本語の翻訳モデル開発に使用した対訳コーパスには，Nowakowski ら[12]が構築したコーパスを用いた．使用したデータ量の合計は 90780 行である．

前処理として，日本語データに関しては MeCab を用いてトークン化を行い，アイヌ語データに関しては主に句読点を正規化し，単語から分離するために，正規表現に基づくカスタムのトークナイザーを用いてトークン化を行った．また，データは訓練データ，検証データ，テストデータに 8:1:1 の割合で分割した．この処理の後の対訳データのまとめは表 1 に表示されている．さらに，NLLB トークナイザーを用いて両方の言語に対し再度トークン化を行った．

表 1　使用したアイヌ語と日本語の対訳データ

データ	行数	トークン数	
		アイヌ語	日本語
訓練	72,624	415,011	539,711
検証	9,078	52,404	67,915
テスト	9,078	51,962	67,600
合計	90,780	519,377	675,226

さらに逆翻訳実験において利用した外部の日本語単言語データには，JParaCrawl [13]コーパスから，アイヌ語と日本語の訓練データの 4 倍（290,496 文）に相当する一部のデータを利用した．具体的には先ず，コーパスに含まれている日本語文を抽出し，重複している文および英数字が 10%以上含まれている文を除去した．次に，文字数に基づき 10 文字ごとのビンに分け，各ビンから，元の訓練データに占める割合を基準としたサンプリングを行うことにより，訓練データにおける長さの分布に近い標本を獲得した．最後に，元の訓練データの日本語側と同様に，MeCab および NLLB トークナイザーでトークン化を行った．

4.2 実験設定

実験では Transformers ライブラリ（v. 4.33）を利用した．全ての実験において，Adafactor オプティマイザーを用いて，初期学習率は2×10^{-5}，全体の学習ステップ数は 50 万，バッチサイズは 32 と設定した．学習率スケジュールは，50 万ステップをかけて，初期学習率から 0 まで線形的に減少するように設定した．5 千ステップごとに，検証データを用いてモデルを評価・保存し，最終的には BLEU スコアが最も高いチェックポイントを採択した．

反復的逆翻訳において，逆翻訳を行う回数は 1〜4 回，サンプリングによって元の対訳コーパスの各文に対して生成する疑似対訳文の個数は 2, 4, 8 個と設定した．なお，外部の日本語単言語データの逆翻訳とテストデータでの評価を行う際は，ビーム幅を 4 としたビーム探索によって対訳文を生成した．

4.3 結果と考察

元の対訳コーパスのみを用いた実験結果を表 2 に示す．表に示す BLEU スコアは，アイヌ語から日本語，日本語からアイヌ語への翻訳結果の平均値である（翻訳方向別の詳細は付録の表 8 を参照）．

表 2　反復的逆翻訳とランダムサンプリングによる疑似対訳文生成に関する実験結果

逆翻訳	疑似対訳文数			
回数	0	2	4	8
0	34.03	—	—	—
1	—	34.28	34.62	34.09
2	—	34.44	34.58	34.71
3	—	34.36	**34.80**	34.28
4	—	34.11	34.63	34.36

逆翻訳を行ったモデルの方がベースラインより高いスコアとなり，逆翻訳の有効性が確認された．反復的逆翻訳に関して，逆翻訳を行う回数を増やすこ

とで, Thanh ら [4]で報告されたようなはっきりしたスコアの上昇傾向は見られなかった. Thanh らの研究では, 反復的に逆翻訳を行うごとに学習全体のステップ数も増えるような設定となっていたため, 反復すればするほど学習量が増えることが精度向上に貢献したという可能性が考えられる.

サンプリングによる疑似対訳文生成に関しては, 逆翻訳回数が 2 回の場合を除き, 生成する疑似対訳文数が 4 個の時に最も高い結果となった. そのことから, サンプリングによって疑似対訳データの多様性をある程度増大させる方法の有効性が確認された. しかし, 生成する疑似対訳文数を 4 個を超えて増加させると, スコアが低下する傾向が見られた. その理由として, 学習ステップの総数を固定したまま学習データの量が増えた結果, 一つ一つのデータを十分に学習しきれていない可能性が考えられる. また疑似対訳文の多様性が増大するにつれ, それらのデータを学習するのに必要なモデルのパラメータの割合が増加すると思われる. しかし人工的に生成した対訳データには多くのエラーが含まれ, 実際のデータ分布から離れているため, 実際のデータにおける性能への効果が限られている可能性がある.

Fadaee ら [14]によると, 元の訓練データを単に複製することでベースラインモデルおよび逆翻訳を用いたモデルよりも良い結果となることが報告された. そこで本研究でも逆翻訳をせず訓練データを 4, 8, 16 倍にオーバーサンプリングする実験を行い, その結果を表 3 に示す.

表 3　ベースラインモデルでオーバーサンプリングを行った場合の実験結果

逆翻訳	オーバーサンプリング倍率			
回数	1[i]	4	8	16
0	34.03	34.10	33.98	**34.16**

オーバーサンプリング率が 4, 16 の場合に, オーバーサンプリングを行わなかったベースラインモデルより少し高いスコアとなった . しかし表 2 と比較すると, オーバーサンプリングよりも逆翻訳の方が精度向上への効果が高いことが確認された.

JParaCrawl データを利用した実験の結果を以下の表 4 に示す. なお, 外部の日本語の単言語データを利用した実験では, 元の訓練データ 1 つにつき 4 つの疑似対訳文を用いた.

表 4　外部の日本語の単言語データを利用した実験結果

逆翻訳	疑似対	目的語		
回数	訳文数	アイヌ語	日本語	平均値
1	4	37.19	**31.36**	**34.28**
2		37.39	30.69	34.04
3		37.33	29.79	33.56
4		**37.43**	30.00	33.72

日本語からアイヌ語の方向ではベースラインの 36.71 より高いスコアとなった. しかし, 外部の日本語単言語データを利用したアイヌ語から日本語の方向では最大のスコアがベースラインの 31.34 に近く, 逆翻訳の回数が増え, JParaCrawl データがより早く訓練データに追加されるにつれ, 結果が悪化する傾向がある. 外部の日本語データの使用によって効果が見られなかった理由として, データのドメインが異なることが考えられる. 評価データを含む元のデータは主に口承文芸の文書から構成されているのに対し, JParaCrawl は Web から獲得された様々な分野の文書を含む. Sennrich ら [3]は, ドメインが異なっても元の訓練データよりも新たな単言語データを用いた方が効果的であることを確認した. しかし, NLLB は日本語データで事前学習されているため, 他ドメインの新たな単言語データを追加することによって得られる効果が限られている可能性がある.

5 おわりに

本研究では, 逆翻訳を用いてアイヌ語と日本語の機械翻訳の精度を改善するために, 反復的逆翻訳とランダムサンプリングによる疑似対訳文生成の手法を使い, さらに外部の日本語データを利用した. その結果, 逆翻訳の有効性は確認されたが, 反復的逆翻訳に関しては明確な精度の上昇傾向が見られなかった. またサンプリングによる疑似対訳文生成では, ある程度有効性が確認されたが, 生成する対訳文数が多すぎる場合には逆に精度の向上度合いが減少した. さらに, 外部の日本語の単言語データを利用した場合よりも元の対訳コーパスのみを使用した場合の方が精度向上に効果的であることが確認された.

[i] オーバーサンプリングを行わない場合である.

謝辞

本研究は JSPS 科研費 JP22K17952 の助成を受けたものです.

参考文献

1. **UNESCO.** Atlas of the world's languages in danger. (オンライン) (引用日: 2024 年 5 月 20 日.) https://unesdoc.unesco.org/ark:/48223/pf0000187026.

2. **文化庁.** 消滅の危機にある言語・方言. (オンライン) (引用日: 2024 年 5 月 20 日.) https://www.bunka.go.jp/seisaku/kokugo_nihongo/kokugo_shisaku/kikigengo/index.html.

3. **Rico Sennrich, Barry Haddow, Alexandra Birch.** Improving Neural Machine Translation Models with Monolingual Data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016.

4. **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少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, 2022.

5. **田中 蒼大郎, 越前谷 博, 荒木 健治.** アイヌ語-日本語ニューラル翻訳における Bilingual BERT による単語埋め込みの適用.令和 5 年度 電気・情報関係学会北海道支部連合大会, 2023.

6. **NLLBTeam.** No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672, 2022.

7. **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 International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages, 2023.

8. **Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andre F. T. Martins, Alexandra Birch.** Marian: Fast Neural Machine Translation in C++. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, 2018.

9. **Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, Trevor Cohn.** Iterative Back-Translation for Neural Machine Translation. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, 2018.

10. **森田 知熙, 秋葉 友良, 塚田 元.** 双方向の逆翻訳を利用したニューラル機械翻訳の教師なし適応の検討.言語処理学会 第 25 回年次大会 発表論文集, 2019.

11. **今村 賢治, 藤田 篤, 隅田 英一郎.** サンプリング生成に基づく複数逆翻訳を用いたニューラル機械翻訳.人工知能学会論文誌, 2019.

12. **Karol Nowakowski, Michal Ptaszynski, Fumito Masui.** A proposal for a unified corpus of the Ainu language.研究報告自然言語処理（NL）, 2018.

13. **NTT Communication Science Laboratories.** JParaCrawl. (オンライン) (引用日: 2024 年 10 月 4 日.) https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/.

14. **Marzieh Fadaee, Arianna Bisazza, Christof Monz.** Data Augmentation for Low-Resource Neural Machine Translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), 2017.

A 付録

表 5 学習時のパラメータ

設定項目	パラメータ名	値
学習率の初期値	learning_rate	2×10^{-5}
ミニバッチサイズ	per_device_train_batch_size / per_device_eval_batch_size	32
トレーニングステップ（合計）	max_steps	500,000
評価頻度（ステップ）	eval_steps	5,000
チェックポイント保存頻度（ステップ）	save_steps	5,000
最適なチェックポイントの指標	metric_for_best_model	BLEU
重み減衰	weight_decay	0.01
オプティマイザー	optim	adafactor

表 6 ランダムサンプリングによる疑似対訳文生成時のパラメータ

設定項目	パラメータ名	値
サンプリング	do_sample	True
ビーム数	num_beams	1[ii]
温度	temperature	1.0
Top-k トークンフィルタリング	top_k	0[iii]
Top-p トークンフィルタリング	top_p	1.0

表 7 外部の日本語データを用いた逆翻訳およびモデル評価時のパラメータ

設定項目	パラメータ名	値
サンプリング	do_sample	False
ビーム数	num_beams	4

表 8 反復的逆翻訳とランダムサンプリングによる疑似対訳文生成に関するそれぞれの方向における実験結果[iv]

逆翻訳回数	目的語	疑似対訳文数 2	4	8
1	アイヌ語	37.00	37.26	36.55
	日本語	31.56	**31.97**	31.62
2	アイヌ語	37.47	37.25	37.55
	日本語	31.41	**31.90**	31.87
3	アイヌ語	37.57	**37.69**	36.90
	日本語	31.14	**31.91**	31.65
4	アイヌ語	37.03	37.46	37.22
	日本語	31.19	31.79	31.50

表 9 ベースラインモデルでオーバーサンプリングを行った場合に関するそれぞれの方向における実験結果

逆翻訳回数	目的語	オーバーサンプリング倍率 1	4	8	16
0	アイヌ語	36.71	36.50	36.47	**36.75**
	日本語	31.34	**31.69**	31.48	31.56

表 10 表 2 で一番良い結果となったモデル[v]を用いてアイヌ語から日本語へ翻訳した結果の例

原文：アイヌ語	1	eminausi a= ominausi kor oka =an .
	2	toan pon seta pirkano a= omap kusu pirkano a= resu .
	3	tanto anakne rera ka isam wa sonno sirpokke .
目的文：日本語 (参照訳)	1	お互いに笑っていました。
	2	あの子犬はよく可愛がられて、よく育てられた。
	3	今日は風もなくて本当に暖かい。
目的文：日本語 (翻訳結果)	1	笑いながら暮らしました。
	2	あの子犬をきれいにかわいがって育てた。
	3	今日は風もなくてとても暖かい

[ii] ビームサーチを行わない場合である.
[iii] Top-k フィルタリングを行わない場合である.
[iv] 太字になっている BLEU スコアは，それぞれの方向（日本語→アイヌ語・アイヌ語→日本語）において，一番良い結果と，一番良い結果との差が 1 未満の結果を示している.
[v] 逆翻訳回数が 3 回で，元の対訳コーパスのみを使ってランダムサンプリングで疑似対訳文を 4 個生成した場合のモデルである．このモデルは次のリポジトリにて公開中である：https://huggingface.co/smilemikan/nllb-ain-jpn-bidir-ibt-best