中間言語を利用したデータ多様化とアンサンブル学習に基づくゼロリソース機械翻訳

Bui Tuan Thanh　秋葉友良　塚田元
豊橋技術科学大学大学院
{bui.tuan.thanh.mg, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

概要

本研究では、対訳データのない（ゼロリソース）言語対のニューラル機械翻訳システムを対象として、中間言語を介して擬似対訳データを構築する手法を提案する。提案手法はソース言語・中間言語と中間言語・ターゲット言語の２つの対訳コーパスを利用し、多様性のあるかつ品質の高いソース・ターゲットの擬似対訳データを構築する。本稿では、中間言語の単言語データを利用することの効果も調べる。実験では、４つの翻訳タスクを用い、提案手法が有効であることを示した。ピボット翻訳手法と比較すると、翻訳性能 (BLEU スコア) は+1.27〜+6.09向上した。

1　はじめに

ニューラル機械翻訳 (Neural machine translation: NMT)[1, 2] は大規模な対訳データを用い、非常に高い性能を実現している。しかし、大規模な対訳データを入手するのは高いコストがかかる。翻訳システムの需要が増えているが、十分な対訳データが揃わない言語対は多くあり、ニューラル機械翻訳の低資源言語問題と呼ばれる。この問題を解決するために、順・逆翻訳で構築できる擬似対訳データを利用してデータを拡張する手法 [3, 4] が提案されている。Nguyen らは擬似対訳データの多様性を高めるために、Data Diversification 手法を提案した [5]。しかし、それらの手法は目的の言語対の対訳データを必要とする。本研究では対訳データのない（ゼロリソース）言語対の翻訳システムを対象とする。

ゼロリソース機械翻訳システムを構築するためのアプローチとして、多言語機械翻訳 [6, 7] とピボットベース手法 [8, 9] の２つがある。多言語機械翻訳アプローチは複数の言語対の対訳コーパスで学習される多言語機械翻訳モデルを用い、学習データに出現していない翻訳方向（ゼロショット翻訳）を行う。このアプローチは直接対訳データを用いなくても翻訳を実現できるが、多くの言語対の翻訳データを必要とする。また、ゼロショット翻訳の精度は高くない。ピボットベース手法は第３言語を中間言語として利用する。このアプローチの代表的な手法はピボット翻訳 [8] である。ピボット翻訳は２つのモデルを用いて２段階で翻訳を行う。この手法は翻訳時間がかかり、エラー伝播の問題がある。そのため、１つのモデルで翻訳できることが望ましい。そこで、中間言語を利用して目的言語対の擬似対訳データを構築する手法が提案されいる。[8, 9]。

本研究では、ピボットベースのアプローチに基づき、中間言語を介した２つの対訳データを利用して目的言語対の擬似対訳データを構築する。擬似対訳データの多様性を高めるために、Data Diversification 手法を適用する。また、より良い擬似対訳データを作成するために、モデルのアンサンブルを適用する。それらの２つの手法の利点を共に活かすために、２つの手法を組み合わせる手法を提案する。さらに中間言語の単言語データを利用する手法も検討する。実験では英語を中間言語として、タイ語 → ベトナム語、ベトナム語 → タイ語、日本語 → ベトナム語とベトナム語 → 日本語の４つの翻訳タスクを用いる。実験結果により、提案手法が有効であることを示す。

2　関連研究

2.1　ピボット翻訳

ピボット翻訳は中間言語を利用する最も単純な手法であり、統計的機械翻訳でよく使用されていた。ソース言語・ターゲット言語の対訳データがないが、中間言語を介したソース言語・中間言語と中間言語・ターゲット言語の２つ対訳コーパスがある場

合に適用可能である。ピボット翻訳はその２つの対訳コーパスを用い、(ソース言語 → 中間言語) 翻訳モデルと (中間言語 → ターゲット言語) 翻訳モデルを学習し、２段階で翻訳を行う。この手法は第１段階目の翻訳が失敗すると、第２段階の翻訳は誤った入力を受け取ることになり、エラーが増幅する。エラー伝播の問題を避けるために、１つのモデルで翻訳できることが望ましい。

2.2 Data Diversification

Nguyen ら [5] は (ソース言語 → ターゲット言語) と (ターゲット言語 → ソース言語) の複数のモデルを用いて複数の擬似データを構築し、そのデータとオリジナルデータを結合して新たな学習データを構築する手法を提案した。この手法は学習データの多様性を高めることで、翻訳性能を改善できた。本研究では、擬似データの多様性を高めるために、この手法を活用する。

2.3 アンサンブル学習

モデルのアンサンブルは、複数のモデルで予測値を計算する [10]。Wang ら [11] は複数のモデルのアンサンブルの出力を用いるデュアル学習フレームワークを提案した。このフレームワークでは学習時に複数の (ソース言語 → ターゲット言語) モデルのアンサンブルでソースデータをターゲット言語に翻訳し、複数の (ターゲット言語 → ソース言語) モデルのアンサンブルで生成したターゲットデータを介してソースデータを復元できる確率を最大化する。我々は擬似データの生成にモデルのアンサンブルを適用する。

3 提案手法

本研究では、豊富なデータがある英語を介した２つの対訳 (src(ソース言語), en) と (en, tgt（ターゲット）) を活用し、ソース・ターゲットの擬似対訳データを構築する手法を提案する。そして、構築した擬似対訳データを用いてモデルを学習する。さらに、中間言語の英語の単言語データも利用する。

ベース手法 先行研究 [8, 9] ではソース言語の擬似データのみまたはターゲット言語の擬似データのみを使用しているのに対し、本研究はその２つの擬似データを使用する。２つの言語対（src, en）と (en, tgt) の対訳データを用いて en2src(英語 → ソース言語) と en2tgt(英語 → ターゲット言語) の翻訳モデルを学習する。en2src モデルで (en, tgt) の英語データをソース言語に翻訳し、ソースの擬似データ src’ を生成する。src’ はターゲットの tgt データに対応しているため、（src’, tgt）の擬似対訳データを構築できる。同様に en2tgt モデルを用いて（src, tgt’）の対訳データを構築する。そのような順・逆翻訳で構築できた擬似対訳データを構築できる（付録. 図１）。そのデータでソース → ターゲットモデルを学習する。英語の単言語データを用いる際、en2src と en2tgt モデルを使用して擬似データ src’ と tgt’ を生成し、（src’, tgt’）の対訳データを構築する。対訳データで構築できた擬似対訳データと単言語データで構築できた擬似対訳データを区別するために、タグを使用する。翻訳する際、対訳データで構築できたデータを示すタグを用いる。

ベース+DD(Data Diversification) Data Diversification 手法を適用することで、擬似対訳データの多様性を高めることができると考える。異なる N 個の en2src モデルと N 個の en2tgt モデルを学習する。ここではモデルのパラメータの初期値をランダムに設定することで、異なるモデルを学習する。それらのモデルを用いて、順・逆翻訳で複数の擬似対訳データを構築できる（付録. 図２）。構築したすべての擬似データを用いて翻訳モデルを学習する。本研究の実験では N=3 とする。英語の単言語のデータを利用する際、複数の (en2src, en2tgt) モデルの組み合わせで擬似対訳データを構築する。

ベース+アンサンブル モデルのアンサンブルを使用することで、多くのモデルの力でより良い擬似データを作成することができると考える。ここでは異なる N 個の en2src モデルと N 個の en2tgt モデルを活用する。すべての en2src モデルのアンサンブルで (en, tgt) の英語データをソース言語に翻訳し、ソースの擬似データ src’ を生成して擬似対訳データを構築する。同様に、すべての en2tgt モデルのアンサンプルで擬似対訳データを構築する（付録. 図３）。この手法は多くのモデルを使用するが、構築される擬似対訳データのサイズはベース手法と同じである。英語の単言語データを用いる際には、すべての en2src モデルのアンサンブルでソース言語に翻訳し、すべての en2tgt モデルのアンサンブルでターゲット言語に翻訳する。

ベース+DD+アンサンブル：本研究は Data Diversification 手法とアンサンブル手法の利点を同時に活用するために、その２つの手法を組み合わせる手法

を提案する。N 個の en2src モデルから $\binom{N}{N-1}$ = N 通りの組み合わせ、N 個の en2tgt モデルから $\binom{N}{N-1}$ = N 通りの組み合わせを作る。それらの組み合わせのアンサンブルを用い、順・逆翻訳で複数の擬似対訳データを構築する（付録. 図 4）。実験では N=3 とする。英語の単言語のデータを利用する際、複数の (en2src のアンサンブル, en2tgt のアンサンブル) の組み合わせで擬似対訳データを構築する。

4 実験

4.1 実験設定

データ 本実験では（タイ語 (th)、ベトナム語 (vi)）と（日本語 (ja)、ベトナム語 (vi)）を対訳文のない言語対とし、英語を中間言語とした。提案手法を th2vi(タイ語 → ベトナム語)、vi2th(ベトナム語 → タイ語)、ja2vi(日本語 → ベトナム語)、(ベトナム語 → 日本語）という 4 つの翻訳タスクで評価した。International Workshop on Spoken Language Translation (IWSLT) の英語・ベトナム語 (IWSLT2015)、英語・日本語（IWSLT2017)、日本語・ベトナム語 (IWSLT2012)、タイ語・英語（TED2020）とタイ語・ベトナム語（TED2020）の対訳コーパスを用いた。英語・ベトナム語は tst2012 を開発データとし、tst2013 をテストデータとする。英語・日本語に対しては dev2010 を開発データとし、tst2015 をテストデータとして使用した。日本語・ベトナム語に対しては、dev2010 を開発データとし、tst2010 をテストデータとして使用する。タイ語・ベトナム語のデータは TED2020 の 1500 対訳文を開発データとし、2000 対訳文をテストデータとした。英語の単言語データは OPUS サイトの TED2013 から入手した。使用しているデータのサイズを表 1 に示す。

表 1 データセット（文数）

翻訳タスク	学習	開発	テスト
en-vi	133,317	1,553	1,268
en-ja	223,108	871	1,194
th-en	157,262	1,000	2,000
ja-vi	–	558	1,225
th-vi	–	1,500	2,000
en	500,000	–	–

前処理 英語文は Moses ツールキットで、ベトナム語文は pyvi ライブラリーの ViTokenizer で、日本語は Mecab[12] でトークン化した。英語文とベトナム語文は Moses の truecaser で処理した。各言語の文は Byte Pair Encoding(BPE) で 16000 サブワードに分割した。

ハイパーパラメータ 本実験は FAIRSEQ[13] を使用した。Transformer アーキテクチャ [14] を用いて機械翻訳モデルを学習した。すべてのモデルは同じハイパーパラメータで学習した。ハイパーパラメータの設定は学習率が 1×10^{-8}、ウォームアップが 4000 ステップ、学習率減衰が逆平方根、ラベル平滑化が 0.1、ドロップアウトが 0.3、重み減衰が 0.0001、損失関数がラベル平滑化クロスエントロピーとした。Adam の最適化アルゴリズムは $\beta_1 = 0.9, \beta_2 = 0.98$ を使用した。モデル学習は 30 エポックで行われた。

4.2 手法比較

本実験で提案手法の有効性を検証するために、以下の手法と比較して評価した。

ピボット翻訳： ソース → 英語モデル、英語 → ターゲットモデルを用いて翻訳を行う。各モデルは提案手法と同じ対訳データから学習した。

多言語機械翻訳のゼロショット翻訳： 第 3 言語のみ使える条件を設定するため、多言語機械翻訳モデルをソース言語 ↔ 英語、英語 ↔ ターゲット言語の対訳データのみで学習される。実験で使用する多言語機械翻訳モデルは（ソース → 英語、英語 → ターゲット）の 2 つの翻訳を行う MNMT-2 方向モデル、（ソース → 英語、英語 → ソース、ターゲット → 英語、英語 → ターゲット）の 4 つの翻訳を行う MNMT-4 方向モデルの 2 種類を構築した。

また、英語の単言語データを使用することの効果を調べるために、次の 3 つの設定を使用した。

・Parallel：擬似対訳データを構築する際、英語を介した 2 つの対訳データの英語文のみ使用する。

・Monolingual：擬似対訳データを構築する際、英語の単言語データのみ使用する。

・Parallel + Monolingual：擬似対訳データを構築する際、英語を介した 2 つの対訳データの英語文と英語の単言語データを使用する。

4.3 実験結果

実験結果を表 2 に示す。本実験での多言語機械翻訳モデルではゼロショット翻訳は全く機能しなかった。その原因は言語対の数が少ないこと、データ量が不足していることが考えられる。提案手法のベース手法で構築した擬似データで学習したモ

表 2 実験結果

手法	th→vi	vi→th	ja→vi	vi→ja
ピボット翻訳	10.16	10.49	11.31	9.85
MNMT-2 方向 (zero-shot)	0.00	0.09	0.13	0.07
MNMT-4 方向 (zero-shot)	0.06	0.28	0.16	1.05
ベース (Parallel)	13.33	9.90	12.45	10.93
＋ **DD** (Parallel)	15.01	11.40	**13.32**	10.89
＋**アンサンブル** (Parallel)	14.84	11.84	12.43	10.90
＋**アンサンブル+DD**(Parallel)	**15.50**	**12.05**	12.61	**10.98**
ベース (Monolingual)	13.46	6.06	10.16	8.15
＋ **DD** (Monolingual)	**14.81**	6.73	**12.15**	10.47
＋**アンサンブル** (Monolingual)	13.55	7.61	10.86	9.71
＋**アンサンブル+DD**(Monolingual)	14.61	**7.95**	11.61	**10.52**
ベース (Parallel + Monolingual)	14.26	10.52	12.37	10.47
＋ **DD** (Parallel + Monolingual)	15.35	11.27	12.54	10.63
＋**アンサンブル** (Parallel + Monolingual)	15.40	**11.82**	12.49	10.91
＋**アンサンブル+DD** (Parallel + Monolingual)	**16.25**	11.47	**12.78**	**11.12**

デルの翻訳はピボット翻訳より精度が高かったが、ベトナム語 → タイ語の方向ではピボット翻訳の方が良かった。単言語データを用いない際、Data Diversification とモデルのアンサンブルを適用することで、ベース手法より良いモデルを学習できた。Data Diversification とモデルのアンサンブルを組み合わせると、翻訳性能が向上した。ただし日本語 → ベトナム語については、モデルのアンサンブルは翻訳性能を少し低下させた。

単言語データのみで擬似データを構築 (Monolingual) すると、データサイズは（Parallel）の約２倍になったが、性能は低かった。ソースとターゲットの両側のデータが全て擬似データであるため、翻訳品質が低いと思われる。また、Data Diversification に関しては、２つの翻訳方向で、Data Diversification とモデルのアンサンブルを組み合わせたものの性能が一番高かった。

（Parallel+Monolingual）では Data Diversification とモデルのアンサンブルを組み合わせたものが、３つの翻訳方向で一番優れた性能を達成した。ベトナム語 → タイ語モデルのみアンサンブル学習だけ用いる方が性能が高った。

これらの結果から、提案手法の有効性は翻訳方向と言語に依存することと考えられるものの、Data Diversification とモデルのアンサンブルを組み合わせる提案手法は一貫して全ての翻訳方向で精度を向上させることがわかった。また、単言語データで構築した対訳データの両側が擬似データであっても、翻訳性能を改善することに利用できることがわかった。

5 おわりに

本論文では対訳文のない言語対の機械翻訳のために、中間言語を介した２つの対訳データを用いてより良い擬似対訳データを構築する手法を提案した。データの多様性を高める Data Diversification 手法とモデルのアンサンブルを組み合わせる提案手法は一貫して全ての翻訳方向の精度を向上させらことを示した。この手法は翻訳モデル自体に複雑な変更を必要とせず、追加のデータも利用しやすい。また、中間言語の単言語データで構築された擬似対訳データを用いてデータを拡張することで性能を改善できた。本研究では４つの翻訳方向で評価したが、今後より多くの翻訳タスクで提案手法を検証したい。実験では 50 万文の単言語データを用いたが、サイズがまだ小さいと考える。単言語データの規模がどのように影響するかも調査したい。また、学習済み多言語機械翻訳を利用することも検討したい。

謝辞

本研究は JSPS 科研費 23K11118 の助成を受けたものです.

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[2] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda, editors, **Proceedings of the 2nd Workshop on Neural Machine Translation and Generation**, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. Data diversification: A simple strategy for neural machine translation, 2020.

[6] Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 268–277, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation, 2017.

[8] A. Gispert and José B. Mariño. Catalan-english statistical machine translation without parallel corpus: Bridging through spanish. 2006.

[9] Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. Phrase-based statistical machine translation with pivot languages. In **Proceedings of the 5th International Workshop on Spoken Language Translation: Papers**, pp. 143–149, Waikiki, Hawaii, October 20-21 2008.

[10] Michael Perrone and Leon Cooper. When networks disagree: Ensemble methods for hybrid neural networks. **Neural networks for speech and image processing**, 08 1993.

[11] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Multi-agent dual learning. In **International Conference on Learning Representations**, 2019.

[12] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

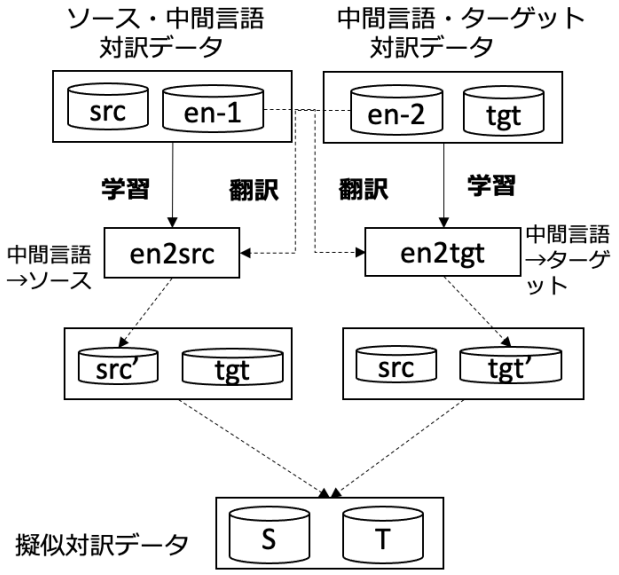

図 1　ベース手法

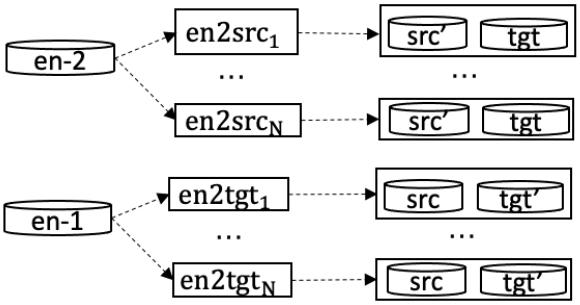

図 2　ベース＋ Data Diversification 手法

A　付録 (Appendix)

この付録では、**3. 提案手法**で説明した手法の図を提供する。ベース手法を図 1 に示す。(src, en-1) はソース・中間言語対訳データであり、(en-2, tgt) は中間言語・ターゲット対訳データである。Data Diversification を適用する手法を図 2 に示す。図 2 には、中間言語のデータ en-2 を複数の en2src モデルでソースに翻訳し、複数のソース言語の擬似データ src' を作成する。モデルのアンサンブルを適用する手法を図 3 に示す。複数のモデルを四角形で囲んだところはモデルのアンサンブルを示す。

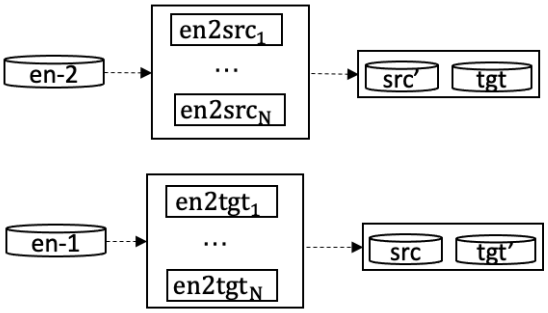

図 3　ベース＋アンサンブル手法

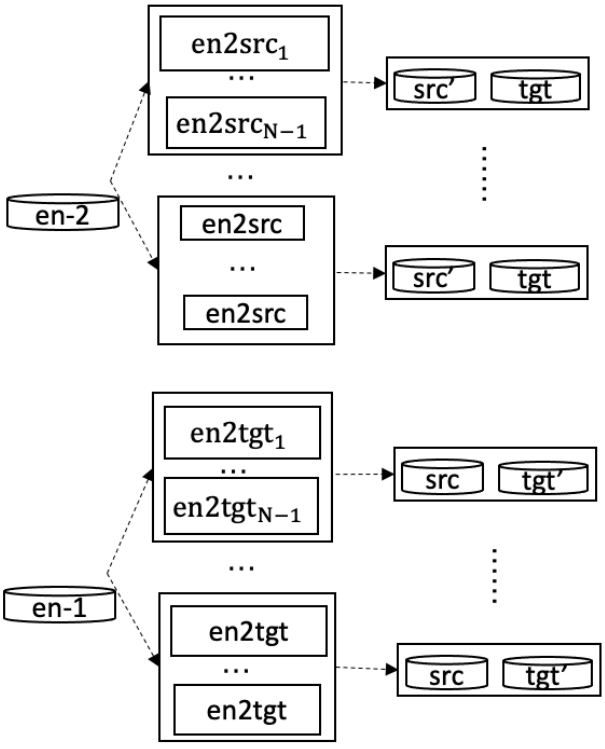

図 4　ベース＋ Data Diversification ＋アンサンブル手法