中間言語を介した２つの対訳コーパスを用いた対訳文のない言語対の NMT の検討

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

ニューラル機械翻訳では高い性能を発揮するために、大規模かつ品質が高い対訳コーパスが必要になる。しかし、対訳コーパスのない言語対が数多くある。本稿では、日本語（日）とベトナム語（越）を対訳文のない言語対として、中間語の英語を介した英日と英越対訳データのみを用いて、日越と越日それぞれの翻訳モデルを学習する手法を提案する。提案手法は、英語との２つの対訳データを使用して、多様性のある日越擬似データを構築する。IWSLT の対訳データを用いた実験により、提案手法は小規模な対訳データで学習された教師あり翻訳モデルやピボット翻訳手法を上回ることを示した。ピボット翻訳手法と比較すると、日越方向の性能 (BLEU スコア) が+1.96、越日方向の性能が+1.75 向上した。

1　はじめに

近年、ニューラルネットワークを用いたニューラル機械翻訳 (Neural Machine Translation: NMT)[1, 2] は従来の統計的機械翻訳の性能を上回り、非常に高い性能を実現している。ニューラル機械翻訳は、翻訳の品質を飛躍的に向上させるために、大規模かつ品質が高い対訳コーパス（学習データ）が必要になる。しかし、そのようなデータを入手するのは困難で、高いコストがかかる。少数言語対においては、利用できる対訳コーパスが少数である低資源言語対の問題、さらには対訳データを利用できないという問題がある。そのような問題に対応するために、低資源言語対の機械翻訳 [3, 4, 5, 6] 及び対訳データのない言語対の機械翻訳の研究が行われている [7, 8, 9]。

対訳文のない言語対の翻訳モデルを学習する最も単純な手法はピボット翻訳 [8] である。ピボット翻訳は２つの翻訳モデルを利用するため、翻訳時間がかかり、エラー伝播の問題がある。Ren らは第３言語を介して低資源言語対の翻訳精度を上げる Triangular Architecture を提案し、その手法が対訳文のない言語対にも応用する見込みがあることを主張している。しかし、その応用の有効性を実験的に示していない。また、単言語コーパスのみを利用し、翻訳モデルを学習する手法 [10, 11, 12] がある。これらの手法では、Sennrich ら [3] の逆翻訳で対訳データを構築する。He ら [7] は逆翻訳がソース側の学習データと評価データのギャップを生じることを示し、順翻訳と逆翻訳で生成された擬似対訳データを組み合わせて利用する手法を提案した。

本稿は、日越を対訳文のない言語対として、英日と英越対訳データのみを用いて、日越と越日それぞれの翻訳モデルを学習する手法を提案する。提案法では英語との２つの対訳データを用いて、順翻訳と逆翻訳で日越の複数の擬似データを構築し、それらの擬似データで翻訳モデルを学習する。実験結果から、この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

ビボット翻訳 [8] は、対訳文のない言語対の翻訳システムを中間言語を介して構築する最も単純な手法である。中間言語とソース言語、中間言語とターゲット言語との対訳データを用いる。この手法では、（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）翻訳モデルを用いて、ソース言語の文をターゲット言語に翻訳する。２段階で翻訳を行うため、翻訳時間がかかる手法である。また、第１段回目の翻訳の失敗が、第２段回の翻訳は誤った入力を翻訳するとこになり、エラーが増幅する。これをエラー伝播と呼ぶ。

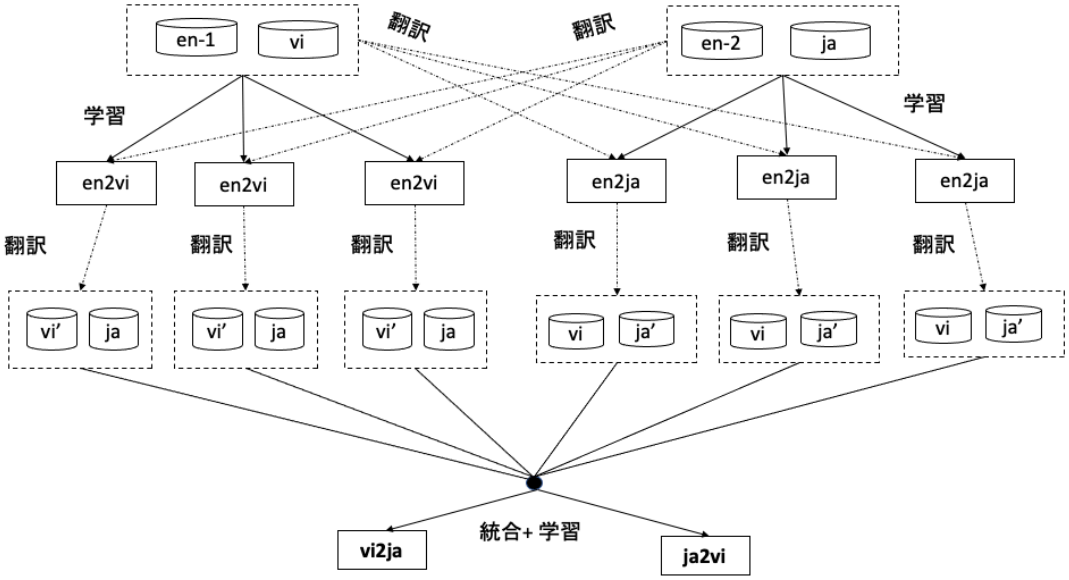

図 1　提案手法の仕組み

2.2　zero-shot 翻訳

Johnson ら [9] は、複数の言語対を用いて 1 つの翻訳モデルを学習することで、直接学習データのない言語対も翻訳できることを示した。Johnson らの研究では 12 の言語対を用いた。この手法では、多くの言語対の対訳データが必要になる。本研究の実験で、2 つの対訳データのみを用いた zero-shot 翻訳は有効ではなかった。

2.3　Triangular Architecture

Ren ら [6] は、豊富な対訳データをもつ中間言語 Y を介して、低資源言語対 X-Z の翻訳制度を上げるのに Triagular Architecture を提案した。X-Y 対訳データを豊富な対訳データとする。Triagular Architecture では、X→Z 翻訳モデルの精度を上げるために、(X,Y) 対訳データと (Y,Z) 対訳データを利用する。X→Z 翻訳モデルを直接学習せずに、Z を潜在変数（latent variable）として X→Y 翻訳モデルを学習する。その後、X→Y 翻訳モデルの学習は低資源言語対 (X,Y) と (Y,Z) それぞれの翻訳モデルを学習することになる。その 2 つのモデルの最適化を双方向 EM アルゴリズムで行う。この手法は単言語のデータを用いる逆翻訳と組み合わせることができる。Ren らは、対訳文のない言語対に応用する見込みがあることを主張しているが、実験的に有効性を示していない。本稿も中間言語を利用しているが、中間言語との対訳データのみを用いて目的の言語対の擬似対訳データを構築する。

2.4　Data Diversification

Nguyen ら [5] は複数の順翻訳モデルと逆翻訳モデルを用いて、それらのモデルで生成された複数の擬似データとオリジナル学習データを統合し、新しい学習データを構築する手法を提案した。Nguyen らの研究によれば、同じモデルの数を用いる際、パラメータの初期化をランダムに実行された複数のモデルはパラメータの初期化を固定した複数のモデルより性能が良くなった。この手法は English-Nepali、Nepali-English、English-Sinhala、Sinhala-English などの低資源言語対に有効性があることが示されている。提案手法では、擬似データの作成に際して Data Diversification の考えを活用する。

3　提案手法

本稿では、対訳データのない日越言語対の翻訳モデルを学習するために、英語を介した 2 つの対訳データのみを用いて日越の擬似データを構築する手法を提案する。提案手法の仕組みを図 1 に示す。ここでは、(en-1、vi) は英越の対訳コーパスであり、(en-2、ja) は英日の対訳コーパスである。 en-1 と en-2 は対になっている必要はない。

最初に、英越と英日の対訳コーパスから、複数の英越翻訳モデル（en2vi）と英日翻訳モデル（en2ja）を学習する。ここでは、seed の値の変更により、各のモデルのパラメータの初期値をランダムに設定す

る。この方法で、異なる翻訳モデルを学習することができる。

次に、学習できたモデルを用いて、擬似対訳コーパスを生成する。具体的には、英日コーパスの英語文（en-2）を英越翻訳モデル（en2vi）でベトナム語に翻訳し、擬似対訳文（vi'、ja）を構築する。同様に、英越コーパスの英語文（en-1）を英日翻訳モデルで日本語に翻訳し、擬似対訳コーパス（vi、ja'）を作成する。同じ文を異なる翻訳モデルで翻訳して擬似対訳文を作ることで、多様性のある擬似データを作ることができる。予備実験では順翻訳または逆翻訳の擬似データのみを用いるよりもそれらのコーパスを統合して学習したほうが性能が高かった。そのため、本稿の提案手法はすべて擬似データを統合して、最終モデルを学習する。

4 実験

4.1 対訳コーパス

本稿では、International Workshop on Sopken Language Translation (IWSLT) の英越対訳コーパス (IWSLT 2015)、英日対訳コーパス (IWSLT 2017) と日越対訳コーパス (IWSLT 2012) を用いる。英越の言語対は tst2012 を開発データとし、tst2013 をテストデータとする。英日言語対に対しては、dev2010 を開発データとし、tst2015 をテストデータとして利用する。日越言語対に対しては、dev2010 を開発データとし、tst2010 をテストデータとして利用する。各言語対の文数を表 1 に示す。

表 1 各言語対の文数

言語対	訓練	開発	テスト
英越	133,317	1,553	1,268
英日	223,108	871	1,194
日越	106,758	558	1,225

英語文は Moses キットでトークン化される。ベトナム語文を pyvi ライブラリの ViTokenizer を用いてトークン化した。日本語は Mecab[13] でトークン化した。また、英語文とベトナム語文は Moses の truecased を用いて処理した。各言語は Byte Pair Encoding (BPE) で 16000 のサブワードに分割した。

4.2 実験条件

fairseq[14] を用いてニューラル機械翻訳システムを構成した。すべてのニューラル機械翻訳モデルは Transformer で学習した。学習の際は、すべてのモデルを同じパイパーパラメータと学習エポック数（=30）を用いて、学習率は 1×10^{-8}、ウォームアップは 4000 ステップ、学習率減衰は逆平方根、ラベル平滑化は 0.1、ドロップアウトは 0.3、重み減衰は 0.0001、損失関数は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では $(\beta_1 = 0.9, \beta_2 = 0.98)$ を使用した。

教師ありモデル: 教師あり翻訳モデルを IWSLT の日越対訳データで学習した。日越対訳データを使用したのは本モデルのみである。以後のモデルは日越対訳データを使用していない。

zero-shot 翻訳: Johnson ら [9] の Many-to-Many 翻訳モデルを用いて zero-shot 翻訳を行った。英日、日英、英越と越英という 4 つの翻訳を可能にする 1 つの翻訳モデルを学習した。モデルを学習する際、各バッチでその 4 つの対訳データの割合を 1:1:1:1 にした。

ピボット翻訳: ピボット翻訳は、英語を中間語として、ソース文を英語に翻訳し、翻訳できた英語文をターゲット言語に翻訳する。たとえば、日越方向翻訳をする際、日本語文を日英翻訳モデルを用いて英語に翻訳する、翻訳できた英語文を英越翻訳モデルでベトナム語に翻訳する。ピボット翻訳に使用したモデルの性能を表 3 に示す。

提案手法: $(en2vi \times k + en2ja \times k)$ は、それぞれ異なる k つの en2vi モデル及び異なる k つの en2ja モデルを用い、擬似対訳コーパスを生成する。$k = 3, 4, 5$ で実験を行った。提案手法の使用されていたモデルを表 4 に示す。

4.3 実験結果

実験結果を表 2 に示す。実験結果により、ピボット翻訳手法は日越両方向で教師ありモデルを上回った。日越と越日それぞれの zero-shot 翻訳はできなかった。その原因は少数の言語対で Many-to-Many 翻訳モデルを学習したと考える。擬似データを用いたことで、日越翻訳性能と越日翻訳性能は向上した。$(en2vi \times 1 + en2ja \times 1)$ はベースラインを比較すると越日方向が＋ 0.83、日越方向が＋ 1.01 向上した。

複数の擬似対訳コーパスを導入すると、越日翻訳性能と日越翻訳性能はさらに改善した。越日翻訳性

能は 13.30 で、日越翻訳性能は 11.65 であった。しかし、en2vi モデルと en2ja モデルをそれぞれ５つ以上のモデルを使用すると、有効に働かなかった。

表 2　実験結果 (dev / test)

手法	vi2ja	ja2vi
教師ありモデル	6.80 / 7.64	8.60 / 9.75
ピボット翻訳	8.55 / 9.90	9.32 / 11.34
zero-shot	1.2 / 1.6	0.4 / 0.4
en2vi $\times 1 + en2ja \times 1$	9.21 / 10.73	10.82 / 12.35
en2vi $\times 3 + en2ja \times 3$	9.20 / 11.21	11.43 / **13.30**
en2vi $\times 4 + en2ja \times 4$	9.70 / **11.65**	11.00 / 12.93
en2vi $\times 5 + en2ja \times 5$	9.45 / 11.45	11.00 / 13.02

表 3　ピボット翻訳に使用したモデル (dev / test)

モデル	翻訳性能
en2vi	26.60 / 30.37
vi2en	25.89 / 29.57
en2ja	12.41 / 13.31
ja2en	7.21 / 9.04

表 4　提案手法で使用したモデル (dev / test)

en2vi	en2ja
26.60 / 30.37	12.41 / 13.29
26.63 / 29.79	12.57 / 13.64
26.54 / 30.11	12.46 / 13.36
26.66 / 29.59	12.22 / 13.48
26.43 / 30.07	12.46 / 13.22

4.4　擬似データの多様性

実験結果から、擬似データの多様性により、日越と越日それぞれの翻訳性能か向上したと考えられる。しかし、多様性ではなく、学習データが増えたから、性能が改善した単に可能性もある。このことを検証するために、($en2vi \times 3 + en2ja \times 3$) の実験を用いて以下の３つのケースで、実験を行った。実験結果を表５に示す。

・**ケース１**：３つの同じ翻訳モデルを使用する。すなわち、これらのモデルで生成された擬似データは全く同じとする。

・**ケース２**：３つのモデルの内、２つのモデルは同じモデルである。

・**ケース３**：３つの異なる翻訳モデルを使用する。これらのモデルで生成された擬似データは異なる。(提案手法)

表 5　擬似データの多様性の有効性 (dev / test)

手法	vi2ja	ja2vi
ケース 1	9.38 / 10.68	10.94 / 12.28
ケース 2	9.28 / 11.12	11.53 / 12.75
ケース 3	9.20 / **11.21**	11.43 / **13.30**

表５の結果により、全く同じ擬似対訳データを用いることより異なる擬似データを使用した方がいいことがわかる。以上の提案手法は多様な擬似データを用いて en2vi モデルと en2ja モデルの翻訳性能を改善することで、日越の擬似対訳コーパスの品質を向上させることができた。

5　おわりに

本稿は、対訳文のない日越言語対の翻訳モデルを学習するために、英日と英越対訳データのみを用いて、日越の擬似対訳データを生成する手法を提案した。英語を介した２つの対訳コーパスで英越と英日それぞれ複数の翻訳モデルを学習した。それらの翻訳モデルを用いて、順翻訳と逆翻訳を組み合わせ、日越の複数の擬似データを生成し、すべての擬似データを統合して、日越と越日それぞれの翻訳モデルを学習した。実験では IWSLT の対訳データを用いて提案手法を検証した。実験結果により、本稿の提案手法は、小規模の対訳データで学習された教師ありモデルおよびピボット翻訳手法を上回ることを示した。さらに、追加実験で、多様性のある擬似データを用いることの有効性を示した。

今後は多様性のある擬似データを構成する手法を用いて、日越の擬似データを生成する英越翻訳モデルと英日翻訳モデルの翻訳性能を改善する予定である。英越翻訳モデルと英日翻訳モデルの翻訳性能が向上すれば、日越の擬似データの品質が改善でき、日越と越日それぞれの翻訳精度はさらに向上させられると考えている。また、提案手法を適用し、他の対訳データのない言語対の翻訳モデルの性能を改善したいと考えている。

謝辞

本研究は JSPS 科研費 18H01062 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. In **Proceedings of the 27th International Conference on Neural Infor-**

mation Processing Systems - Volume 2, NIPS'14, p. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[2] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In **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 **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. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[6] Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma. Triangular architecture for rare language translation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 56–65, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. Bridging the data gap between training and inference for unsupervised neural machine translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**.

[8] DE GISPERT A. Catalan-english statistical machine translation without parallel corpus : Bridging through spanish. **Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 2006**, pp. 65–68, 2006.

[9] 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. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 339–351, 2017.

[10] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In **International Conference on Learning Representations (ICLR)**, 2018.

[11] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 46–55, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In **Proceedings of the Sixth International Conference on Learning Representations**, April 2018.

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

[14] 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.