多言語事前学習モデルを前提とした非英語間ピボット翻訳の特徴調査

今村 賢治　　隅田英一郎
国立研究開発法人 情報通信研究機構
{kenji.imamura,eiichiro.sumita}@nict.go.jp

概要

多言語翻訳モデルは、複数の言語を一つのモデルで扱えるようにしたものであるが、学習していない言語対（ゼロショット翻訳）の翻訳品質は非常に悪い。一方、ピボット翻訳は英語等のピボット言語を介して原言語→ピボット→目的言語の翻訳を行う方法で、原言語と目的言語の直接対訳がなくても機械翻訳を行うことができる。

本稿では、多言語モデルを用いてピボット翻訳を行い、直接翻訳と比較する。また、逆翻訳で生成した疑似対訳で多言語モデルをファインチューニングすることで、直接対訳がなくても翻訳品質を向上できることを示す。

1　はじめに

多言語ニューラルネットワークモデルは、複数の言語を一つのモデルで学習させたもので、機械翻訳や言語横断処理に有効である。処理対象に類似した言語（たとえば、同じ言語属のもの）もリソースとして活用することで、低リソースでも処理対象言語の精度を向上させられるという利点がある。[1)]機械翻訳の場合、原言語と目的言語の組み合わせで翻訳を行うため、言語特化モデルでは組み合わせ数のモデルを用意しなければならない。多言語モデルを用いると、さまざまな言語対を一つのモデルで扱えるため、管理が容易となる。そのため、対訳コーパスで事前学習したエンコーダ・デコーダモデルもいくつか公開されている。

たとえば、OPUS-100 コーパスを事前学習した多言語翻訳モデル [1][2)] が公開されている。これは、100 言語と英語間の翻訳モデル（いわゆる英語中心 English-Centric モデル）である。M2M-100 モデル [2][3)] の対象も 100 言語であるが、英語以外の対訳コーパスを追加することで、2,200 方向の事前学習を行っている。なお、mBART [3, 4] もエンコーダ・デコーダ型の事前学習モデルだが、これは単言語コーパスだけで訓練されている。

100 言語を翻訳対象とした場合、のべ 9,900 方向の翻訳を行う可能性があるが、多言語翻訳モデルを用いても、対訳コーパスで学習されていない言語対（ゼロショット翻訳 [5] と呼ばれている）の翻訳品質は非常に悪く、実用に適さない場合が多い。

対訳コーパスの入手が困難な言語対で、ある程度精度のよい翻訳を実現する方法として、ピボット翻訳 [6] がある。これは、原言語と目的言語の間にピボット言語を設け、ピボットを介して原言語→ピボット言語→目的言語を実現する（図 1(a)）。ピボットには、対訳コーパスが豊富な言語が有利なため、英語が用いられることが多い。ピボット翻訳は統計翻訳でよく利用されていたが、ニューラル機械翻訳にも適用可能である。多言語事前学習モデルにピボット翻訳を適用することで、対訳コーパスが存在しない言語間（ゼロリソース言語。主に非英語間）の翻訳であっても、1 つのモデルで実用的な翻訳が実現できる。

本稿では、ゼロリソース言語対に対して、多言語事前学習モデルを用いたピボット翻訳を適用し、翻訳品質の測定と、逆翻訳を併用した品質向上法について議論する。

2　本稿の多言語事前学習モデル

今回、CC-100 コーパス [7, 8] と、OPUS-100 コーパス [9, 10] または CCAligned v1 コーパス [11] がカバーする 103 言語に対応した英語中心モデルを新た

1）　高リソース条件では、多言語モデルより一つの言語に特化した言語特化モデルの方が、一般的には精度は高い（多言語の呪い (curse of multilinguality) と呼ばれている）。

2）　https://github.com/bzhangGo/zero/tree/master/docs/multilingual_laln_lalt#pretrained-multilingual-models-many-to-many

3）　https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100

表 1　コーパスサイズ

コーパス	文数 訓練	開発	テスト
ALT	18,088	1,000	1,018
ASPEC-JC	669,923	2,090	2,107

に訓練した。CC-100 は単言語コーパス、OPUS-100 と CCAligned は対訳コーパスである。いずれも Web のクロールデータを基にしている。

構築方法は以下のとおりである。

1. まず、[12] の方法を踏襲し、mBART-50 [3] モデルの単語埋込を、CC-100 コーパスがカバーする 109 言語に拡張した。拡張部分はランダム初期化している。
2. 次に、CC-100 コーパスを用いて、上記モデルに対してノイズ除去の追加訓練した。この訓練は、mBART-50 の訓練と同じである。
3. その後、OPUS-100 と CCAligned コーパスのうち、英語 ↔ 外国語対訳を用いて、モデルを追加訓練した。言語対によって、コーパスサイズに大きな差があるため、温度サンプリング [13] を行い、コーパスサイズが大きな言語対はダウンサンプリング、小さな言語対はアップサンプリングしながら訓練した（温度係数 $T = 0.7$）。

作成されたモデルは、基本的には mBART-50 と同じ構造なので、エンコーダ、デコーダはそれぞれ 12 層、埋込 1,024 次元、FFN4,096 次元、16 ヘッド、単語埋込は 25 万語である。なお、mBART-50 は、原言語と目的言語を言語タグという形で指定するため、本モデルも翻訳時に原言語、目的言語が明確になっている必要がある。

本モデルの訓練コーパスサイズ（一部）と、3 節で用いる ALT コーパスの言語と英語間の翻訳品質を付録に示す。

3　翻訳実験

本稿では、外国語と日本語間翻訳について実験を行う。

3.1　実験設定

3.1.1　コーパス

今回は、ゼロリソース言語と、直接対訳が存在した場合を比較するため、以下の対訳コーパスを利用した。コーパスサイズを表 1 に示す。なお、訓練

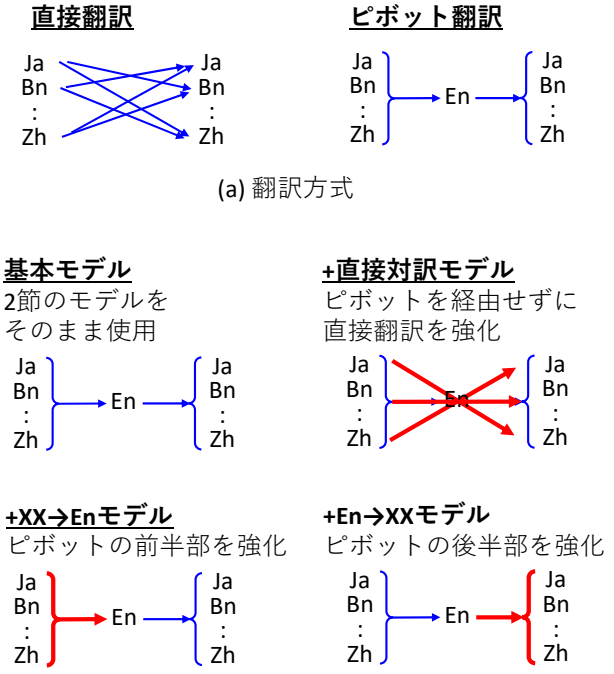

図 1　翻訳方式とモデル

セットは、著しく長さが異なる対訳を取り除いたあとのサイズである。

低リソース言語の実験では、Asian Language Treebank (ALT) Parallel コーパス [14][4] を使用した。これは、英語 (En)、日本語 (Ja) の他、ベンガル語 (Bn)、インドネシア語 (Id)、クメール語 (Km)、ラオ語 (Lo)、マレー語 (Ms)、ミャンマー語 (My)、タイ語 (Th)、タガログ語 (Tl)、ベトナム語 (Vi)、中国語簡体字 (Zh) をカバーする多言語コーパスである。同じ英語 Wikinews の文を各言語に翻訳しているため、英語以外の言語間でも対訳となっているのが特徴である。実験は、日本語と英語以外の外国語間の翻訳について行う。

中リソース言語の実験では、ASPEC-JC[15][5] を使用する。これは日本語 (Ja) と中国語 (Zh) の対訳コーパスで、科学技術文献を基にしている。対応する英語はない。主として、逆翻訳の有効性を確認するために使用する。

3.1.2　比較方式・翻訳システム

本稿では、直接翻訳とピボット翻訳の比較を行う（図 1(a)）。その際、使用するモデルは、2 節の多言語事前学習モデルをそのまま使用する場合を基本

4）https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
5）https://jipsti.jst.go.jp/aspec/

モデルとし、それを対訳コーパスでファインチューニングしたモデルの翻訳結果と比較する（図 1(b)）。ファインチューニングは、以下の対訳コーパスを用いて行った。

- **+直接対訳:**
  外国語 ↔ 日本語対訳でファインチューニングした場合。ピボットを経由してないため、翻訳方式は直接翻訳となる。[6)]
- **+XX→En:**
  外国語 → 英語対訳でファインチューニングした場合。ピボット翻訳の前半部を強化した状態となる。なお、ピボット翻訳の後半部は基本モデルを使用する。
- **+En→XX:**
  英語 → 外国語対訳でファインチューニングした場合。ピボット翻訳の後半部を強化した状態となる。ピボット翻訳の前半部は基本モデルを使用する。

ピボット翻訳を用いると、ゼロリソース言語対でも機械翻訳を行うことができるので、直接対訳コーパス（人手翻訳）以外にも、目的言語のコーパスをピボット経由で逆翻訳した疑似対訳を用いてファインチューニングすることもできる。本稿では、人手翻訳と逆翻訳で作成した疑似対訳の比較も行う。なお、原言語から作成した疑似対訳は本稿では順翻訳と呼ぶ。逆翻訳は単語サンプリング [17, 18] を使って、目的言語 1 文に対して 1 つの疑似原文を生成した。順翻訳は、翻訳品質が直接最終翻訳品質に影響するため、1 ベスト訳を使用した。

3.1.3 その他の実験設定

ファインチューニングとテスト時のハイパーパラメータを表 2 に示す。

評価は sacreBLEU [19] で行った。ミャンマー語など、トークナイザーが対応していない言語は BLEU [20] で評価するのが不適切な場合もあるため、トークナイザー非依存の評価方法である ChrF [21] の評価結果を併記する。

3.2 実験結果

ALT コーパスによる実験結果を表 3、ASPEC-JC コーパスによる実験結果を表 4 に示す。なお、ALT コーパスの結果は外国語 ↔ 日本語の結果の平均値を示している。

6) ファインチューニングしていない言語は catastrophic forgetting 現象 [16] により翻訳品質が著しく低下するため、ピボット翻訳は適用できない。

表 2 ハイパーパラメータ一覧

種別	値
ファインチューニング	温度サンプリング [13]: $T = 0.7$, Loss: label_smoothed_cross_entropy=0.1, Dropout: 0.3, Warmup: 約 1 エポック, LR: 0.00008, inverse_sqrt, Early Stopping: 10 エポック, Batch サイズ: 8K トークン, Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-6}$)
テスト	ビーム幅: 10

3.2.1 ピボット翻訳 vs. 直接翻訳

ALT コーパスも、ASPEC-JC コーパスも、言語方向によらず、基本モデルで直接翻訳した結果 (No. 1) は非常に低い値となっている。これはゼロショット翻訳となっているためである。しかし、同じモデルでピボット翻訳を使う (No. 2) と、12 ポイント以上の BLEU スコアで翻訳することができ、ゼロリソース言語でもある程度の翻訳品質が確保できる。

直接対訳でファインチューニングすると（No. 5a）、No. 2 に比べも ALT コーパスで 3 ポイント以上、ASPEC-JC コーパスでは 18 ポイント以上、BLEU スコアが向上する。ALT コーパスは 1.8 万文しかファインチューニングに使用していないので、翻訳品質向上を目指すなら、できる限り直接対訳を整備すべきである。

3.2.2 ピボット翻訳の前半部と後半部

表 3 の No. 3a、4a はそれぞれ基本モデルの一部の言語を対訳コーパスでファインチューニングしたものである。No. 3a はピボットの前半部、No. 4a は後半部を強化したことに相当する。

No. 2 と No.3a は、大きな差異がないが、No. 4a は BLEU スコアが 3 ポイント程度向上している。つまり、ピボット翻訳の前半部と後半部では、後半部を強化した方が、効率的に品質を向上させることができた。これは、基本モデルが Web から取得したコーパスで学習されているおり、ドメインが異なっているためで、ピボット後半をドメイン適応させることで品質向上したものと考えられる。

表 3　ALT コーパスの翻訳結果。太字はモデル・翻訳方式で最高品質を表す。

No.	モデル	翻訳方式	Ja → XX 平均 BLEU	Ja → XX 平均 ChrF2	XX → Ja 平均 BLEU	XX → Ja 平均 ChrF2	備考
1	基本モデル	直接翻訳	0.2	6.3	0.1	0.8	ゼロショット翻訳
2	基本モデル	ピボット	12.6	40.4	17.3	26.9	ベースライン
3a	+XX → En (人手翻訳)	ピボット	12.6	40.0	18.8	28.3	
4a	+En → XX (人手翻訳)	ピボット	**15.7**	**45.6**	**21.4**	30.2	
5a	+直接対訳 (人手翻訳)	直接翻訳	15.4	44.8	21.2	**30.3**	上限

表 4　ASPEC-JC コーパスの翻訳結果。太字はモデル・翻訳方式で最高品質を表す。

No.	モデル	翻訳方式	Ja → Zh BLEU	Ja → Zh ChrF2	Zh → Ja BLEU	Zh → Ja ChrF2	備考
1	基本モデル	直接翻訳	0.0	0.0	0.1	0.2	ゼロショット翻訳
2	基本モデル	ピボット	19.4	17.6	12.0	21.8	ベースライン
3b	+ XX → En (順翻訳)	ピボット	19.6	17.7	12.4	22.2	
4b	+ En → XX (逆翻訳)	ピボット	26.8	23.2	19.2	27.8	
5b	+ 直接対訳 (逆翻訳)	直接翻訳	30.8	26.2	24.3	32.9	
5a	+ 直接対訳 (人手翻訳)	直接翻訳	**37.6**	**32.0**	**33.4**	**41.6**	上限

3.2.3　機械翻訳によるデータ拡張

表 3、表 4 の No. 3a, 4a, 5a は、人手翻訳による対訳コーパスでファインチューニングしたモデル、No. 3b, 4b, 5b は、機械翻訳（逆翻訳または順翻訳）で作成した疑似対訳でファインチューニングしたモデルによる結果である。

ピボット翻訳のベースライン (No. 2) と比べると、ピボットの前半部を疑似対訳で強化 (No. 3b) しても、翻訳品質はほとんど変わらない。

一方、逆翻訳で作成した疑似対訳でピボットの後半部を強化 (No. 4b) したり、直接対訳でファインチューニングすると (No. 5b)、ベースラインより大幅に翻訳品質を向上させることができる。人手翻訳 (No. 5a) の翻訳品質には到達できないが、逆翻訳は目的言語の単言語コーパスがあれば可能である。単言語コーパスが豊富にあるならば、ゼロリソース言語対でも、（疑似）直接対訳をピボット逆翻訳で作成しファインチューニングすれば、効率的に翻訳品質を向上させることができると考えられる。

4　まとめ

ピボット翻訳はゼロリソース言語対でも翻訳が可能である。目的言語の単言語コーパスがあれば、逆翻訳により疑似対訳コーパスを生成することができるので、ゼロリソース状態でも翻訳品質を向上させることができる。翻訳品質そのものは直接翻訳の方が高いため、実用システムを目指すなら、直接対訳コーパスを充実させるべきであるが、直接翻訳が実現できない場合でも、これに準じる性能を実現できるピボット翻訳は有用である。さらに、機械翻訳を新しい言語対に拡張する際、翻訳結果を確認したり、逆翻訳で作成した疑似対訳を後編集することで直接対訳作成の補助にも利用できる。

ピボット翻訳と直接翻訳を適切に使い分けながら、多言語化を進める予定である。

謝辞

本件は、総務省の「ICT 重点技術の研究開発プロジェクト (JPMI00316)」における「多言語翻訳技術の高度化に関する研究開発」による委託を受けて実施した研究開発による成果です。

参考文献

[1] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. **arXiv e-print**, 2004.11867, 2020.

[2] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. **arXiv e-print**, 2010.11125, 2020.

[3] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 726–742, 2020.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan

Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[5] Melvin Johnson, Mike Schuster, Quoc 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, No. 0, pp. 339–351, 2017.

[6] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In **Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference**, pp. 484–491, Rochester, New York, April 2007.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[8] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[9] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)**, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[11] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5960–5969, Online, November 2020. Association for Computational Linguistics.

[12] Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. Extending multilingual BERT to low-resource languages. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 2649–2656, Online, November 2020. Association for Computational Linguistics.

[13] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. **arXiv e-print**, 1907.05019, 2019.

[14] Hammam Riza, Gunarso Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. Introduction of the asian language treebank. In **Oriental COCOSDA**, 2016.

[15] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[16] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. **arXiv preprint**, 2013.

[17] Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In **Proceedings of the 2nd Workshop on Neural Machine Translation and Generation**, pp. 55–63, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[18] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[19] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[21] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

表 5　基本モデルの訓練コーパスと英語間翻訳品質

XX 言語	訓練コーパスサイズ (文) CC-100	OPUS-100	CCAligned	En → XX BLEU	ChrF2	XX → En BLEU	ChrF2
日本語 (Ja)	393M	1.0M	15.0M	26.0	36.0	26.2	57.2
ベンガル語 (Bn)	54M	1.0M	3.5M	9.5 †	44.5	28.2	56.8
インドネシア語 (Id)	969M	1.0M	15.7M	41.8	67.5	43.0	67.6
クメール語 (Km)	6.6M	0.1M	0.4M	5.2 ‡	47.6	27.0	55.2
ラオ語 (Lo)	2.6M	-	0.2M	2.9 †	24.9	6.3	27.7
マレー語 (Ms)	66M	1.0M	5.4M	44.1	69.1	44.5	68.3
ミャンマー語 (My)	2.0M	0.02M	0.3M	0.0 ‡	36.2	19.3	49.1
タイ語 (Th)	295M	1.0M	10.7M	13.0 ‡	48.2	26.9	56.4
タガログ語 (Tl)	27M	-	6.6M	30.7	59.1	39.2	63.3
ベトナム語 (Vi)	939M	1.0M	12.4M	39.7	57.8	36.0	61.9
中国語簡体字 (Zh)	169M	1.0M	15.2M	35.0	31.0	24.8	56.2

A　基本モデルの英語 ↔ 外国語間翻訳品質

本稿で対象とした言語について、基本モデルの訓練コーパスサイズと、英語間翻訳品質を表 5 に示す。CC-100 は単言語コーパスの文数、OPUS-100 と CCAligned は英語 ↔ 外国語間の対訳文数である。

翻訳品質は、ALT コーパステストセットについて、基本モデルで翻訳したもである。翻訳方式は直接翻訳を使用した。なお、表中の ‡ は、sacreBLEU のトークナイザーが非対応かつスペース区切りが（ほぼ）ない言語を表す。† はトークナイザーが非対応だが、スペース区切りがある言語を表す。