JParaCrawl v4.0: クラウドソーシングを併用した大規模対訳コーパスの構築

森下 睦, 帖佐 克己, 永田 昌明
[1]NTT コミュニケーション科学基礎研究所
makoto.morishita@ntt.com

概要

現在の機械翻訳モデルは主に対訳コーパスを用いて学習されており、その翻訳精度は対訳コーパスの質と量に大きく依存している。本稿では、新たにウェブをクロールし日英対訳文を抽出することで大規模日英対訳コーパスを構築し、翻訳精度の底上げを狙う。なおこの際クラウドソーシングを活用して対訳文が存在するウェブサイトを発見することで、効率的な対訳文収集を目指す。今回ウェブから収集した対訳文と以前作成した日英対訳コーパス JParaCrawl v3.0 を合わせることで、合計 4400 万文を超える日英最大規模の対訳コーパスを作成することに成功した。実験により、新たな対訳コーパスを用いて学習した翻訳モデルが様々な分野で高い翻訳精度を発揮することを示す。なお、今回作成した対訳コーパスを JParaCrawl v4.0 と名付け、我々のウェブサイト上で研究目的利用に限り無償公開する予定である。[1)]

1 はじめに

現在のニューラル機械翻訳モデルは、主に対訳コーパスを用いた教師ありの手法 [2, 3, 4, 5] で学習されている。また大規模言語モデル (LLM) の事前学習時に対訳コーパスを使用することで、多言語文の理解・生成性能が向上することが知られており [6]、対訳コーパスの重要性はさらに増している。これらのモデル学習時、対訳コーパスの質と量が翻訳・言語理解性能に大きな影響を与えること知られているが、大規模対訳コーパスが一般に公開されている言語対は限られている。例えば、独英などの言語対ではすでに数億文の対訳文が公開されているものの、日英ではまだ同程度のものは存在せず、モデル学習時に大きな問題となっている。そのため、本稿ではさらに大規模なウェブベースの日英対訳コーパスを構築する。現在、日英で最大規模の対訳コーパスの 1 つは約 2100 万文の対訳文を含む JParaCrawl v3.0 [7] であり、これはウェブを大規模にクロールし対訳文を自動的に抽出することで構築されている。本コーパスは欧州言語対と比較すると小規模であり、2022 年を最後に更新が止まっているため、最新の情報を含んでいない。そのため、本研究ではウェブを全面的に再クロールし、対訳文を抽出することで JParaCrawl コーパスを拡大/更新する。本研究では、この際クラウドワーカーから日英対訳文が存在するウェブサイトの報告を受けることで、より効率良く対訳文を抽出することを目指す。また、新たに作成した対訳コーパスを用いて、英日および日英の機械翻訳の精度がどのように向上するかを実験的に示す。本研究で作成した対訳コーパスは JParaCrawl v4.0 と名付け、今後の研究のために我々のウェブサイト[2)]で公開する予定である。

2 関連研究

対訳コーパスは様々な文書から対訳文を抽出することで作成されることが多い。例えば、欧州議会の議事録から作成された Europarl [8]、国連の翻訳文書から作成された UN 対訳コーパス [9] などがある。これらの文書は、通常プロの翻訳者が翻訳しており、文書 ID などのメタデータを持っていることもあるので、容易に対訳文を抽出することができる。しかし、一般にメタデータが整った状態で公開されている対訳文書は多くない。

近年では、ウェブから対訳文を抽出する手法も多く提案されている。ウェブ上には 2 言語以上で書かれたウェブサイトが多数存在し、こういったウェブサイトから対訳文を抽出する。ウェブ上には、様々な言語や分野の対訳文が存在しており、大規模な対

1） 本稿は JParaCrawl v3.0 発表時の原稿 [1] をもとに加筆修正したものである。

2） http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

表 1 JParaCrawl コーパスに含まれる重複を取り除いた対訳文数および英語側単語数

バージョン	文数	単語数
v1.0	4,817,172	125,216,523
v2.0	8,809,771	234,393,978
v3.0	21,481,513	502,445,763
v4.0	44,240,792	1,112,091,871

表 2 収集対象ウェブサイト数および対訳文抽出成功ウェブサイト数

列挙手法	収集対象数	抽出成功数
CommonCrawl	50,000	17,675
クラウドソーシング	20,600	17,462
既存コーパス解析	24,000	21,673

訳コーパスを作成するための有望な情報源である。ウェブから対訳文を抽出する初期の研究としては、大規模な分散システムを構築し対訳文を抽出したもの [10]、Common Crawl[3] から対訳文を抽出したもの [11] などがある。また、多言語文埋め込みを用いた対訳文対応手法を用いて、Wikipedia や Common Crawl から大規模な多言語対訳コーパスを作成する研究も報告されている [12, 13]。

また、ParaCrawl プロジェクトはヨーロッパ言語の大規模な対訳コーパスをウェブから継続的に作成している [14]。我々は以前この活動にヒントを得て、大規模な対訳コーパスが存在しない日英向けの大規模な対訳コーパスを作成した [15, 7]。このコーパスは JParaCrawl と名付けられ、2100 万文を超える対訳文を含む日英における最大規模の対訳コーパスとなっている。しかし、JParaCrawl コーパスは、独英などの欧州言語対と比較するとまだ小規模であり、これをもとにした翻訳モデルの精度も欧州言語対と比較すると低精度である。ゆえに、さらに大きな日英対訳コーパスの作成が求められている。本研究では、ウェブを新たにクロールし対訳文をさらに抽出することで、JParaCrawl コーパスをさらに拡張することを目指す。

3 JParaCrawl v4.0

本研究では、ウェブから対訳文を抽出することで大規模な対訳コーパスを構築する。この際の基本的な収集作業は以下の 4 ステップとなる。

3.1 対訳文を含むウェブサイトの発見

本研究ではまず対訳文が存在すると思われるウェブサイト (対訳ウェブサイト) をリストアップし、これらのサイトから対訳文を抽出することを考える。本工程は成果物となる対訳コーパスの質と量に大きく影響する。そのため、JParaCrawl v3.0 以前で使用していた CommonCrawl 解析による手法に加え、クラウドソーシングを活用した手法、既存対訳コーパスの解析による手法を併用して対訳ウェブサイトリストを作成した。各手法により列挙されたクロール対象ドメイン数を表 2 に示す。

CommonCrawl 解析 CommonCrawl 上のテキストデータを言語判定ツール `CLD2`[4] によって解析し、各ドメインの言語別データ量を得る。その後、英語と日本語が同量程度含まれるウェブサイトには対訳文が存在する可能性があるという仮説に基づき、クロール対象ウェブサイトを列挙する。本研究では、2021 年 9 月から 2023 年 6 月までに公開された Common Crawl のテキストアーカイブデータ 12 セット (合計 104TB) を分析対象とし、ウェブサイトの規模が大きく、英語と日本語の文章が同程度であるウェブサイトを列挙した。2021 年 8 月以前に公開されたデータについては、JParaCrawl v3.0 作成時に既に分析済みであるため除外した。なお、本手順には ParaCrawl プロジェクトが提供する `extractor`[5] を使用した。JParaCrawl v2.0 までは本手法のみを使用した。JParaCrawl v3.0 では本手法を主力とし、試験的に一部次項のクラウドソーシングを実施した。

クラウドソーシングの活用 Morishita ら [16] は、クラウドワーカーはこれまでの経験をもとに対訳ウェブサイトを手動で効率的に列挙できるという仮説を立て、ドメイン適応のための対訳文収集時にクラウドソーシングを活用する手法を提案した。また、JParaCrawl v3.0 作成時は、試験的にクラウドソーシング由来の対訳ウェブサイトを少量クロールし、CommonCrawl 由来のウェブサイトより対訳文抽出効率が高いことを確認した [7]。上記の結果を受け今回の対訳コーパス構築時も、より大規模にクラウドソーシングを活用することで効率的に対訳文を抽出できると考えた。本研究では、対象ドメインを絞らずクラウドワーカーに日英対訳が存在するウェブサイトの報告を依頼し、列挙されたウェブサ

3) https://commoncrawl.org/

4) https://github.com/CLD2Owners/cld2

5) https://github.com/paracrawl/extractor

イトをクロール対象とした。[6)]

既存対訳コーパスの解析　本研究では、これまで対訳文が大量に抽出できたウェブサイトは、時間の経過により有用な対訳文が増加している可能性があるという仮説を立てた。前回作成した JParaCrawl v3.0 には、各対訳文がどのドメインから得られたものかがメタデータとして付与されている。そのため、JParaCrawl v3.0 から多く対訳文が抽出できた上位ドメインを列挙し、再クロール対象した。

3.2　ウェブサイトのクロール

次に、前節で列挙されたウェブサイト全体をクロールする。本研究では、`Heritrix`[7)] を用いて、各ウェブサイトに対して最大 48 時間のクローリングを行った。この際 JParaCrawl v3.0 にならい、プレーンテキストに加え、Word、PDF ファイルについてもクロール対象とした。

3.3　対訳文抽出

次に、クロールされたウェブサイトから対訳文を抽出する。本手順には、ParaCrawl プロジェクトが提供する `Bitextor`[8)] を日本語に対応させ使用した。対訳文書と対訳文の対応付けには、機械翻訳を用いた対応付けツール `bleualign` [17] を使用した。このツールでは、まず機械翻訳を用いて日本語文を英文に翻訳し、BLEU スコアを最大化する日英の文ペアを発見する。この際、日英翻訳には JParaCrawl v3.0 で学習した Transformer ベースのニューラル機械翻訳（NMT）モデルを使用した。

3.4　ノイズ除去

最後の手順として、正しく対応付けられていない、翻訳が不正確など、学習時のノイズとなる文対をフィルタリングする。本手順には、`Bicleaner`[9)] [18] を使用した。

3.5　作成結果

上記手順により得られたクリーンな対訳文と JParaCrawl v3.0 を結合し、重複文を削除した。これにより 4400 万文以上を含む新しい大規模日英対訳コーパス JParaCrawl v4.0 の構築に成功した。表 1 に、これまでの JParaCrawl および今回作成した v4.0 の重複を削除した対訳文数および英語側単語数を示す。また、表 2 に、クリーンな対訳文を 1 文以上含んでいた対訳ウェブサイト数 (抽出成功数) を示す。CommonCrawl 由来の収集対象と比較して、クラウドソーシングにより列挙された対訳ウェブサイトは、対訳文抽出の成功率が高くより効率的にクロールできていることがわかる。なお、クラウドソーシングにより列挙されたウェブサイトのうち、CommonCrawl により得られたサイトと重複していたものは約 6%であり、ほとんどが CommonCrawl からは発見できない対訳ウェブサイトであったことも特筆に値する。

4　実験

本節では、新たに作成した JParaCrawl v4.0 の翻訳精度への影響を確認するために、NMT モデルを学習し様々なテストセットでその精度を評価した。以降では、使用したテストセットの詳細およびモデル学習時の設定について述べる。

4.1　実験設定

4.1.1　テストセット

様々な分野で NMT モデルの精度を評価するために、19 種類のテストセットでモデルを評価する。付録表 4 に使用するテストセットの分野および統計情報を示す。これらには、以前の JParaCrawl 発表時に報告した ASPEC [19] (科学技術論文), JESC [20] (映画字幕), KFTT [21] (Wikipedia 記事), TED (tst2015) [22] (講演) などが含まれる。さらに本実験では、WMT22、WMT23 の General 翻訳タスク [23, 24] で使用されたテストセットを追加した。これらは、汎用的な機械翻訳精度を測ることを目的として設計され、ニュース文、対話文、SNS 上の文書、インターネット通販サイトの文書等が含まれている。なおシェアードタスク用テストセットの中には、特定の翻訳方向 (英→日など) で使用することを前提としたものもあるが、参考までに英日、日英の両方向で評価した。また付録表 5 に示すように、いくつかのコーパスには学習データが付属しているものがある。比較のため、これらの学習データで分野別モデルを学習した際のスコアを付録の表 6 に示す。

6)　なお先行研究ではワーカーの挙動に応じて追加報酬を与えているが、本研究では手続きの簡略化のため、全てのワーカーを時給制で雇用した。
7)　https://github.com/internetarchive/heritrix3
8)　https://github.com/bitextor/bitextor
9)　https://github.com/bitextor/bicleaner

表 3 JParaCrawl で学習した翻訳モデルの自動評価値 (BLEU / COMET)。JParaCrawl モデルのうち最高スコアのものを太字で示す。E-J は英日、J-E は日英翻訳を想定して作成されたテストセットであることを示す。

	英日翻訳				日英翻訳			
テストセット	v1.0	v2.0	v3.0	v4.0	v1.0	v2.0	v3.0	v4.0
ASPEC	24.7 / 87.7	26.5 / 88.4	**26.8** / 88.5	26.6 / **88.6**	18.3 / 81.6	19.7 / 82.4	20.8 / 82.8	**21.4 / 83.2**
JESC	**6.6** / 72.1	6.5 / 72.3	6.5 / **72.6**	6.2 / 71.8	7.0 / 66.0	7.5 / 67.6	**8.4** / 68.1	8.2 / **68.2**
KFTT	17.1 / 80.8	**18.9** / 82.4	18.1 / 82.4	18.1 / **82.5**	13.7 / 72.8	16.2 / 74.8	17.0 / 74.6	**17.8 / 75.3**
TED (tst2015)	11.5 / 76.8	12.6 / 78.0	13.1 / 78.9	**13.3 / 79.4**	11.0 / 74.4	11.9 / 75.3	**12.0 / 75.8**	11.9 / 75.6
BSD	12.4 / 83.7	13.5 / 84.4	13.9 / 85.4	**15.2 / 86.0**	17.4 / 79.4	19.6 / 81.0	19.9 / 81.4	**20.0 / 81.7**
WMT20 News E-J	20.7 / 83.6	21.9 / 84.7	23.5 / 85.4	**24.2 / 85.8**	21.3 / 84.0	23.3 / 84.9	**23.9 / 85.4**	23.7 / **85.4**
WMT20 News J-E	20.1 / 85.8	22.8 / 87.0	23.5 / **87.7**	**24.4** / 87.6	19.2 / 78.3	21.0 / 79.3	21.9 / 80.1	**22.9 / 80.4**
WMT21 News E-J	21.1 / 82.4	21.8 / 83.9	25.0 / 84.9	**26.0 / 85.5**	21.9 / 83.8	23.1 / 84.7	**24.3 / 85.2**	24.2 / 85.1
WMT21 News J-E	19.6 / 84.2	21.5 / 85.4	22.4 / 85.7	**23.1 / 86.1**	18.1 / 75.7	20.7 / 77.0	21.3 / 77.6	**23.2 / 78.2**
WMT22 General E-J	20.3 / 81.8	21.1 / 83.3	23.9 / 85.4	**24.9 / 86.0**	21.9 / 84.2	23.4 / 84.9	23.8 / 85.4	**24.3 / 85.6**
WMT22 General J-E	16.1 / 84.7	17.7 / 85.8	18.6 / 86.5	**19.3 / 86.9**	18.4 / 78.1	20.5 / 79.4	21.3 / 80.1	**22.5 / 80.3**
WMT23 General E-J	17.6 / 79.7	18.3 / 81.2	20.5 / 82.4	**21.0 / 82.8**	19.0 / 81.5	19.6 / 82.2	20.9 / 82.6	**20.9 / 83.0**
WMT23 General J-E	16.9 / 85.5	19.2 / 86.9	19.8 / 87.3	**20.4 / 87.7**	17.6 / 77.2	18.8 / 78.4	20.2 / 79.2	**21.4 / 79.8**
WMT19 Robust E-J	12.4 / 72.7	12.5 / 73.3	14.4 / 76.6	14.2 / **76.9**	15.6 / 76.1	16.8 / 77.7	**17.3** / 78.2	16.4 / **78.5**
WMT19 Robust J-E	11.5 / 77.1	12.3 / 78.4	12.8 / 79.2	**13.7 / 80.4**	16.0 / 70.9	17.2 / 73.8	17.7 / 74.8	**18.3 / 75.2**
WMT20 Robust Set1 E-J	15.2 / 64.6	15.8 / 64.8	**18.7 / 67.0**	18.5 / 66.4	20.0 / 70.4	20.6 / 71.6	21.6 / 72.7	**22.7 / 73.0**
WMT20 Robust Set2 E-J	12.7 / 72.6	13.0 / 72.9	**14.8** / 75.3	**14.8 / 75.7**	16.4 / 75.8	17.4 / 77.3	**17.9** / 78.0	17.4 / **78.1**
WMT20 Robust Set2 J-E	7.9 / 76.9	8.2 / 78.1	**8.6 / 78.9**	8.4 / **78.9**	12.0 / 67.8	12.6 / 69.3	14.0 / 69.9	**14.5 / 70.4**
IWSLT21 Simul. E-J Dev	12.5 / 81.4	13.3 / 83.1	14.5 / 82.6	**15.0 / 84.0**	12.9 / 79.3	14.3 / 80.9	14.5 / 81.1	**15.4 / 81.4**

4.1.2 学習設定

翻訳モデルの学習には `fairseq` [25] を用い、small、base、big の 3 つの異なる大きさの Transformer [5] モデルを学習した。付録表 7 に詳細な学習設定を示す。以前の JParaCrawl の報告に基づき、TED (tst2015) では small モデルを、KFTT では base モデルを、その他のテストセットでは big モデルを使用した。自動評価値として `sacreBLEU` [26] を用いた BLEU スコア [27]、`Unbabel/wmt22-comet-da` モデルを用いた COMET スコア [28] を報告する。なお、以前の実験との整合性を保つために、すべてのテストセットを NFKC 正規化した。詳細な実験設定については付録 A に示す。

4.2 実験結果

表 3 に様々なテストセットにおける BLEU スコアおよび COMET スコアを示す。以降では BLEU スコアより人手評価との相関が高いとされる COMET スコアを中心に議論する [29]。JParaCrawl v4.0 で学習したモデルは、19 のテストセットのうち英日では 16、日英では 17 のテストセットで以前の JParaCrawl を上回る精度を達成した。科学技術論文、ニュース、SNS などの様々な分野において、新しい JParaCrawl が NMT モデルの精度を押し上げている。WMT20、21 ニュース翻訳タスクの精度向上ついては、以前報告されているように、ウェブページから近年のニュース文に関連した対訳文を抽出できたことによるものだと考えられる。[7]。また、WMT22、23 General 翻訳タスクについても精度が向上している。複数領域の対訳文から構成されている General 翻訳タスクテストセットと様々な領域を含むウェブデータの相性が良いことが要因の一つとして考えられる。

5 おわりに

本研究では、これまでの大規模日英対訳コーパス JParaCrawl をさらに拡張した JParaCrawl v4.0 を作成した。本対訳コーパスは、従来の CommonCrawl 解析に加え、クラウドソーシングの活用および既存対訳コーパスの解析を行い、対象となったウェブサイトをクロール、対訳文抽出を行うことで作成した。新たな JParaCrawl v4.0 は 4400 万以上の対訳文を含んでおり、これは JParaCrawl v3.0 の倍以上の大きさである。本対訳コーパスを用いることで、様々な分野の翻訳精度が向上することを実験的に確認した。今後の課題としては、継続的な JParaCrawl の更新や、より優れた対訳ウェブサイト検出方法/対訳文抽出手法/フィルタリング手法の提案などが挙げられる。また、より大規模な対訳モデルで本コーパスを学習した際の翻訳精度や、LLM で本コーパスを活用した際の効果は計測できておらず、今後取り組んでいきたいと考えている。なお、本研究で作成した JParaCrawl v4.0 は我々のウェブサイトで研究目的に限り無償で公開する予定である。

参考文献

[1] 森下睦, 帖佐克己, 鈴木潤, 永田昌明. JParaCrawl v3.0: 大規模日英対訳コーパス. 言語処理学会第 28 回年次大会 (NLP2022), 2022.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 3104–3112, 2014.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[4] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1412–1421, 2015.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS), pp. 6000–6010, 2017.

[6] Rohan Anil, et al. Palm 2 technical report. arXiv preprint arXiv:2305.10403, 2023.

[7] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6704–6710, 2022.

[8] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the Machine Translation Summit X, pp. 79–86, 2005.

[9] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pp. 3530–3534, 2016.

[10] Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 1101–1109, 2010.

[11] Jason R Smith, Herve Saint-Amand, M Plamada, P Koehn, C Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1374–1383, 2013.

[12] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. arXiv preprint arXiv:1907.05791, 2019.

[13] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. CCMatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944, 2019.

[14] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4555–4567, 2020.

[15] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 3603–3609, 2020.

[16] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. Domain adaptation of machine translation with crowdworkers. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 606–618, 2022.

[17] Rico Sennrich and Martin Volk. Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA), pp. 175–182, 2011.

[18] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 955–962, 2018.

[19] 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 10th International Conference on Language Resources and Evaluation (LREC), 2016.

[20] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English Subtitle Corpus. arXiv preprint arXiv:1710.10639, 2017.

[21] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[22] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: web inventory of transcribed and translated talks. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 261–268, 2012.

[23] Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 1–45, 2022.

[24] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Proceedings of the Eighth Conference on Machine Translation, pp. 1–42, 2023.

[25] 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: Human Language Technologies (NAACL HLT), pp. 48–53, 2019.

[26] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 186–191, 2018.

[27] 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 (ACL), pp. 311–318, 2002.

[28] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702, 2020.

[29] Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Proceedings of the Eighth Conference on Machine Translation, pp. 578–628, 2023.

[30] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation (WAT), pp. 54–61, 2019.

[31] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 1–55, 2020.

[32] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In Proceedings of the 6th Conference on Machine Translation (WMT), pp. 1–93, 2021.

[33] Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. Findings of the first shared task on machine translation robustness. In Proceedings of the 4th Conference on Machine Translation (WMT), 2019.

[34] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. Findings of the WMT 2020 shared task on machine translation robustness. In Proceedings of the 5th Conference on Machine Translation (WMT), pp. 76–91, 2020.

[35] Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT), pp. 1–29, 2021.

[36] Graham Neubig. Forest-to-string SMT for asian language translation: NAIST at WAT2014. In Proceedings of the 1st Workshop on Asian Translation (WAT), pp. 20–25, 2014.

[37] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, Vol. 15, pp. 1929–1958, 2014.

[39] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on Machine Learning (ICML), Vol. 28, pp. 1310–1318, 2013.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In Proceedings of CVPR, pp. 2818–2826, 2016.

[41] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In Proceedings of the 3rd Conference on Machine Translation (WMT), pp. 1–9, 2018.

[42] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

[43] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 66–71, 2018.

表 4　テストセットに含まれる分野、対訳文数および英語側単語数

テストセット	分野	文数	単語数
ASPEC [19]	科学技術論文	1,812	39,573
JESC [20]	映画字幕	2,000	13,617
KFTT [21]	Wikipedia 記事	1,160	22,063
TED (tst2015) [22]	TED Talk	1,194	20,367
BSD [30]	対話	2,120	19,619
WMT20 News E-J [31]	ニュース	1,000	22,141
WMT20 News J-E [31]	ニュース	993	24,423
WMT21 News E-J [32]	ニュース	1,000	23,305
WMT21 News J-E [32]	ニュース	1,005	24,771
WMT22 General E-J [23]	汎用	2,037	34,037
WMT22 General J-E [23]	汎用	2,008	29,100
WMT23 General E-J [24]	汎用	2,074	35,086
WMT23 General J-E [24]	汎用	1,992	33,542
WMT19 Robust E-J [33]	Reddit	1,392	19,988
WMT19 Robust J-E [33]	Reddit	1,111	13,390
WMT20 Robust Set1 E-J [34]	Wikipedia コメント	1,100	29,419
WMT20 Robust Set2 E-J [34]	Reddit	1,376	20,011
WMT20 Robust Set2 J-E [34]	Reddit	997	15,866
IWSLT21 Simul. Trans. E-J Dev [35]	TED Talk	1,442	20,677

表 5　分野別学習データに含まれる文数および英語側単語数。ASPEC は本来約 300 万文の学習データを含んでいるが、先行研究に基づき先頭 200 万文のみを学習に使用した [36]。

データ	文数	単語数
ASPEC	3,008,500	68,929,413
JESC	2,797,388	19,339,040
KFTT	440,288	9,737,715
TED	223,108	3,877,868

表 6　分野別コーパスのみで学習した翻訳モデルの自動評価値 (BLEU / COMET)

データ	英日	日英
ASPEC	44.3 / 90.4	28.7 / 82.5
JESC	14.5 / 76.0	17.8 / 70.9
KFTT	31.8 / 85.2	23.4 / 78.5
TED	11.1 / 76.0	13.7 / 73.1

A　詳細な実験設定

表 4 にテストセットに含まれる分野、対訳文数および英語側単語数を示す。また、表 5 に分野別学習データに含まれる文数および英語側単語数を示す。なお、ASPEC は本来約 300 万文の学習データを含んでいるが、先行研究に基づき先頭 200 万文のみを学習に使用した [36]。分野別学習データのみで学習した翻訳モデルの自動評価値を表 6 に示す。

JParaCrawl v4.0 を用いた翻訳モデル学習時の前処理として、対訳コーパスを `sentencepiece` [43] を用いて unigram 確率をもとにサブワード単位に分割した。この際、語彙数は 64,000、character coverage は 0.9999 とした [10]。また、未知文字については Byte-fallback により UTF-8 のバイト列の組み合わせによる表現を行った。翻訳モデルを学習する際の詳細なハイパーパラメータを表 7 に示す。

10）JParaCrawl v3.0 作成時は語彙数を 32,000 としたが、コーパスの拡張に伴い各単語埋め込みの学習が十分に行われると考え語彙数を増加させた。

表 7　ハイパーパラメータの一覧

共通設定	
Architecture	Transformer [5]
Enc-Dec Layers	6
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$) [37]
Learning Rate Schedule	Inverse square root decay
Warmup Steps	4,000
Max Learning Rate	0.001
Dropout	0.3 [38]
Gradient Clipping	1.0 [39]
Label Smoothing	$\epsilon_{ls} = 0.1$ [40]
Mini-batch Size	320,000 tokens [41]
Number of Updates	40,000 steps (v4.0), 36,000 steps (v3.0), 24,000 steps (v1.0, v2.0)
Averaging	100 ステップごとにモデルを保存し、最終 8 チェックポイントの平均を用いる
Beam Size	6 (文長による正規化付き) [42]
small 設定	
Attention Heads	4
Word-embedding Dimension	512
Feed-forward dimension	1,024
base 設定	
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
big 設定	
Attention heads	16
Word-embedding dimension	1,024
Feed-forward dimension	4,096