JParaCrawl Chinese v2.0: クラウドソーシングを用いた日中対訳コーパスの構築

永田 昌明　帖佐 克己　安田 宜仁
NTT コミュニケーション科学基礎研究所
{masaaki.nagata,katsuki.chousa,norihito.yasuda}@ntt.com

概要

我々はクラウドソーシングを使って、日中対訳 web サイトのトップページ URL の対を約 1 万件収集し、約 460 万文対の日中対訳コーパスを作成した。まずトップページ URL を起点としてそのドメインをクロールし、次に 16 万語対の日中対訳辞書を用いて文書対応と文対応を行い、最後に別途用意した 120 万文対の高品質な日中対訳文対から作成した対訳コーパスフィルタを用いてフィルタリングを行った。我々の日中対訳コーパス 460 万文対は、既存の日中対訳コーパス CCMatrix(1,240 万文対)[1] に比べ、大きさはの約 3 分の 1 であるが、翻訳精度は同等であり、クラウドソーシングの有効性を示せた。

1　はじめに

従来のエンコーダデコーダモデルだけでなく、大規模言語モデル (Large Language Model, LLM) でも、対訳データは機械翻訳において重要な役割を果たす。Briakou ら [2] は、PaLM の訓練データの分析から、対訳データが訓練データに偶然混入していることが LLM で翻訳ができる原因であることを示した。

10B 前後の小さな LLM は翻訳精度が低いが、Xu ら [3] は、大量の単言語データで LLM をファインチューンした後に高品質かつ少量の対訳データで LLM をファインチューンする ALMA と呼ばれる方法を提案し、13B の LLM で GPT-3 と同等の翻訳精度を達成した。Guo ら [4] および近藤ら [5] は、大規模な対訳データを用いて継続事前訓練してから高品質な対訳データでファインチューンすると、ALMA を上回る翻訳精度を達成できることを示した。

本報告では、日本語と中国語の対訳文対を web から収集する方法について議論する。日本語と中国語の翻訳は、話者の数や経済規模の観点から非英語言語対の中で最も重要な課題と言える。我々は、対訳を含む web サイトの URL をクラウドソーシングにより収集することの有効性を示す。

森下ら [6] は、日英対訳コーパスの収集において、対象領域を制限せずに対訳 web サイトにおける各言語のトップページ URL 対をクラウドワーカを使って収集する方法の有効性を示した。

本報告ではこの方法を日本語と中国語に適用し、CommonCrawl を分析して収集した web サイトよりもクラウドソーシングで収集した web サイトの方が少ないクロール量で多くの対訳文対が得られることを示す。

なおクラウドソーシングを用いて作成した 460 万文対の日中対訳コーパス JParaCrawl Chinese v2.0 は、研究目的利用に限り無償で公開している。[1)]

2　関連研究

2.1　Web からの対訳データ収集

Web をマイニングして対訳データを収集する研究 [7, 8] は 2000 年頃から始まった。これらは局所的マイニングと大局的マイニングに大別できる。

局所的 (local) マイニングは、階層的 (hierarchical) マイニングと呼ぶこともある。Web の階層構造に基づいて、まず対訳文書を含んでいる web サイトを探索し、そのサイトの中で対訳文書対を探索し、対訳文書の中で対訳文対を探索する。大局的 (global) マイニングは、web 全体を構造を持たない巨大な文の集合とみなし、多言語文埋め込みに基づく文の類似度からある文の異なる言語への翻訳を探索する。前者の代表例は ParaCrawl[9]、後者の代表例は CCMatrix[1] である。

Web 全体を対象とする大局的マイニングには膨大な計算資源が必要なので、我々は、日英対訳データを収集した JParaCrawl[10] と同様に、日本語と中国

1） https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

語を対象とする局所的マイニングを採用する。

局所的マイニングにおいて、文書対応や文対応の手法については多くの従来研究があるが、対訳文書を含む web サイトを見つける方法については従来研究が少ない。ParaCrawl[9] では、Common Crawl のアーカイブを対象として、一つのサイトのすべてのページに言語識別器を適用し、収集対象とする言語対のテキストを同じくらい含むサイトを探す。CCAligned [11] では、Common Crawl のアーカイブを分析対象として、URL に含まれる言語を識別可能な文字列を手掛かりとして、互いに翻訳になっている URL 対を探す。

2.2 日中対訳コーパス

最も広く使われている日中対訳コーパスは、日本の科学技術論文の概要を人手で中国語に翻訳した ASPEC-JC (68 万文対) [12] である。また最大の日中対訳コーパスは、中国語と日本語の公開特許公報の対から収集した JPO-NICT 中日対訳コーパス (1.3 億文対) である[2]である。

Wikipedia から収集された日中対訳データには、LinguaTools-WikiTitles v2014 (170 万文対), WikiMatrix (130 万文対) [13], and Wikipedia Chinese-Japanese Parallel Corpus (13 万文対) [14] などがある。映画字幕から収集された日中対訳データには、 OpenSubtitles v2018 (110 万文対) [15] がある。

Web から収集されて公開されている最大の日中対訳コーパスは CCMatrix [1](約 1200 万文対) である。同様に web から収集されて公開されている JParaCrawl Chinese v1.0 [16] は、83,000 文対である。The Asian Language Treebank[17] は、英語の Wikinews を日本語や中国語を含むアジア言語に翻訳したもので、約 2 万文対ある。

WCC-JC 3.0 [18] は、映画やテレビの字幕、歌詞、ニュース記事などを web から収集した約 300 万文対の日中対訳コーパスである。論文の著者にメールを送れば研究目的に限定して入手可能である。

3 方法論

3.1 対訳 web サイト探索

対訳データ収集手順は基本的に ParaCrawl [9] と同じである。Bitextor[3] を用いて、web クロール、文書対応、文対応、対訳コーパスフィルタリングを行う。

ParaCrawl では、Common Crawl のアーカイブを解析してクロールすべき web サイトを決める。言語識別のために CLD2 [4] を各 web ページに適用し、収集対象となる言語対のテキストを同じくらい含む web サイトを抽出する。そして Heritrix [5] を用いて各 web サイトをクロールする。

我々は、2021 年 9 月から 2023 年 6 月までの 12 個の Common Crawl アーカイブに言語識別器 CLD2 を適用し、日本語と中国語のテキストを同じぐらい含むサイトをテキストの総量でソートして、上位 4 万件 web サイトを抽出した。この手続きには、ParaCrawl プロジェクトの Extractor[6] を使用した。

また我々は、クラウドワーカに対して、互いに翻訳になっている web ページを含む web サイトを探し、そのサイトの日本語と中国語のトップページの URL の対を報告するように依頼した。

Common Crawl 解析で得られた web サイトとクラウドソーシングで得られた web サイトの両方について、各サイトを最大 48 時間クロールし、HTML と Word と PDF を収集した。

3.2 文書対応と文対応

Bitextor の文書対応と文対応では、意味的等価性を判定する手段として機械翻訳に基づく方法と対訳辞書に基づく方法を選択できる。我々は、外部の言語資源への依存を最小限にするために対訳辞書に基づく方法を選択した。

Bitextor の対訳辞書に基づく方法では、対訳辞書を使って得られる特徴と HTML の構造から得られる特徴を用いて文書の類似度を計算する。文対応は Hunalign [19][7] を用いる。

文書対応と文対応では、日本語の単語分割に mecab[8] を使用し、中国語の単語分割に jieba[9] を使用した。対訳辞書には、EDR 日中対訳辞書 (533,957 語対) [20][10] を使用した。

文書対応と文対応の計算量を削減するために、EDR 日中対訳辞書の日本語と中国語の見出し語に単語分割を適用し、1 対 1 対応となる 157,900 語対を抽出した。これに日本語の漢字と中国語の簡体字

2) https://alaginrc.nict.go.jp/jpo-outline.html
3) https://github.com/bitextor/bitextor
4) https://github.com/CLD2Owners/cld2
5) https://github.com/internetarchive/heritrix3
6) https://github.com/paracrawl/extractor
7) http://mokk.bme.hu/resources/hunalign/
8) https://taku910.github.io/mecab/
9) https://github.com/fxsjy/jieba
10) https://www2.nict.go.jp/ipp/EDR/JPN/J_HotNews.html

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　対訳 web サイト探索法の比較

対訳サイト探索法	URL 数	エラー数	クロール数	対訳存在数 (率)	文対数
Common Crawl 解析	40,000	19,878	20,122	5,483 (0.272)	2,786,467
クラウドソーシング	11,184	168	11,016	8,204 (0.745)	4,602,328

の対応表を加えた約 16 万語対を使用した。

3.3　対訳コーパスフィルタリング

ParaCrawl プロジェクトには、Bicleaner と Bicleaner AI の 2 つの対訳コーパスフィルタがある。Bicleaner [21] は、単語翻訳確率と統計的言語モデルを特徴として、入力された文対が対訳か否かを判別する random-forest 分類器を訓練する。Bicleaner AI [22] は、訓練済み多言語モデルを用いて二値分類器を訓練する。どちらの方法も分類器を訓練するために高品質な対訳データを必要とする。

我々は、外部言語資源への依存を最小限にし、かつ、計算量を削減するために、Bicleaner[11] を使用した。日本語の単語分割に mecab、中国語の単語分割に pkuseg [23][12] を使用し、日本語と中国語の単語対応に AWESOME-align [24] を使用して、高品質な日中対訳文データから単語翻訳確率を求めた。

対訳コーパスフィルタの訓練には、旅行会話、辞書例文、文学作品、新聞記事などから構成される社内の日中対訳データ (120 万文対) を使用した。この中では旅行会話基本表現集 (BTEC, 約 50 万文対)[25] が最大で、その次に多いのは辞書例文 (約 26 万文対) である。

クロールデータから文書対応と文対応を経て得られた対訳文対から、日中 Bicleaner モデルを使ってスコア 0.5 以上の文対を抽出した。さらに LaBSE[26] を使って cosine 距離が 0.7 以上の文対を抽出した。

表 1 に、Common Crawl 解析とクラウドソーシングにより得られた web サイトについて、クロールに成功したサイト数、少なくとも 1 つ以上の対訳文対が得られたサイト数、収集された対訳文対の総数を示す。

クロールに成功したサイト数に対する 1 つ以上の対訳文対が得られたサイト数の割合は、Common Crawl 解析では 27.2%なのに対して、クラウドソーシングでは 74.5%と非常に高い。最終的に得られた対訳文対数は、Common Crawl 解析が 280 万文対、クラウドソーシングが 460 万文対であり、Common Crawl 解析に比べてクラウドソーシングは少ないクロール量で多くの対訳データを得られることが分かる。

11) https://github.com/bitextor/bicleaner
12) https://github.com/lancopku/pkuseg-python

4　翻訳実験

4.1　データセット

表 2　実験で使用した日中対訳データセット

	train	dev	test
CCMatrix	12,403,136		
WikiTitles	1,661,273		
WikiMatrix	1,325,674		
OpenSubtitles2018	1,091,295		
Crowdsourcing (ours)	4,643,867		
news-commentary-v18		1,625	
Asian Language Treebank		1,000	1,000
ASPEC-JC			2,107
FLORES-200		997	1,012
NTREX-128			1,997
bitext_cj			1,000
WMT2023j			992
total	2,1125,245	3,622	8,126

クラウドソーシングで得られた対訳文対の品質を評価するために、日中翻訳と中日翻訳の精度を調べた。表 2 に実験で使用した日中対訳データセットを示す。

公開されている 100 万文対以上の日中対訳という基準で、比較対象として CCMatrix [1], WikiTitles [27], WikiMatrix [13], OpenSubtitles2018 [15] を選んだ。WikiTitles, WikiMatrix, OpenSubtitles2018 を一つ (wt-wm-os) にまとめ、ccmatrix, wt-wm-os, crowdsourcing の 3 つの翻訳モデルを作成した。

開発セットとして、news-commentary-v18 (1,677 文対)[13], Asian Language Treebank の dev (1,000 文対)[14], FLORES-200 の dev (997 sentence pairs)[15] を使用した。

公開されているテストセットとして、Asian Language Treebank の test (1,000 文対), ASPEC-JC の test (2107 文対), FLORES-200 の devtest (1012 文対), NTREX-128 (1997 文対) [28] を使用した。これら以外に、社内の日中対訳データからランダムに 1000 文対を選択したテストセット (bitext_cj) と、WMT-2023[29] の日英翻訳のテストセットのうち、news

13) https://data.statmt.org/news-commentary/v18.1/
14) https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html
15) https://github.com/facebookresearch/flores

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

(495 文) と質問応答 (497 文) を日本語から中国語へ翻訳したもの (wmt2023j) をテストセットとして用意した。テストセットのソース言語は、ASPEC-JC と wmt2023j は日本語、bitext_cj は日本語と中国語、その他は英語である。

4.2 実験条件

表 3 ハイパーパラメタ

architecture	transformer_wmt_en_de_big
enc-dec layers	6
optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
learning rate schedule	inverse square root decay
warmup steps	4,000
max learning rate	0.001
dropout	0.3
gradient clip	0.1
batch size	1M tokens
max number of updates	60K steps
validate interval updates	1K steps
patience	5

翻訳ソフトウェアとして fairseq [30]、翻訳モデルとして transformer big [31] を使用した。表 3 に Transformer ハイパーパラメタを示す。トークナイザは sentencepiece [32] を使用し、日本語と中国語の語彙をそれぞれ 32k とした。翻訳精度は、sacreBLEU [33, 34] と COMET (wmt22-comet-da) [35] で評価した。

4.3 翻訳精度

表 4 日本語から中国語への翻訳

	ccmatrix		wt-wm-os		crowdsourcing	
test set	bleu	comet	bleu	comet	bleu	comet
ALT	34.4	0.856	18.9	0.779	35.8	0.847
ASPEC	35.8	0.856	17.6	0.767	37.8	0.862
flores200	29.5	0.860	16.0	0.776	33.8	0.863
ntrex128	25.2	0.815	14.2	0.735	25.4	0.806
bitext_cj	22.3	0.808	11.9	0.739	23.8	0.812
wmt2023j	23.9	0.801	11.7	0.713	32.1	0.824
average	28.5	0.833	15.1	0.751	**31.5**	**0.836**

表 5 中国語から日本語への翻訳

	ccmatrix		wt-wm-os		crowdsourcing	
testiest	bleu	comet	bleu	comet	bleu	comet
ALT	24.1	0.886	15.9	0.817	22.6	0.872
ASPEC	29.7	0.896	19.9	0.834	29.9	0.897
flores200	26.2	0.887	14.4	0.795	26.6	0.881
ntrex128	18.8	0.859	11.8	0.775	17.5	0.844
bitext_cj	17.6	0.833	8.8	0.755	16.5	0.833
wmt2023j	24.9	0.874	14.1	0.782	23.6	0.878
average	**23.6**	**0.872**	14.15	0.793	22.8	0.867

表 4 に日中翻訳の精度、表 5 に中日翻訳の精度を示す。3 つの翻訳モデル ccmatrix, wt-wm-os, crowdsourcing の中では、ccmatrix と crowdsourcing の翻訳精度が同じぐらで、wt-wm-os は精度が低い。ccmatrix と crowdsourcing では、日本語から中国語への翻訳では crowdsourcing が高く、中国語から日本語への翻訳では ccmatrix が高い。

5 議論

Crowdsourcing (460 万文対) は、 CCMatrix (1240 万文対) の 1/3 の量でほぼ同じ翻訳精度である。従ってクラウドソーシングで得られた対訳は高品質であると言ってよい。しかし、Crowdsoucing は、CCMatrix と比べると日中翻訳の精度は高いが中日翻訳の精度が低い。これはクラウドソーシングが日本において日本人により行われたため、収集された web サイトの多くが日本語を中国語へ翻訳したもので、中国語の多様性が低いためと思われる。

翻訳実験ではクラウドソーシングによる日中対訳 (460 万文対) のみを使用したが、Common Crawl 解析による日中対訳 (280 万文対) を加えれば、中国語の多様性が高まって中日翻訳の精度が向上すると予想される。これとは別に、中国語をソース言語として日本語へ人手で翻訳されたテストセットがないので、そもそも中日翻訳の自動評価はあまり信頼できないという問題がある。

6 おわりに

本報告では、クラウドソーシングを用いて互いに翻訳になっている web ページを含む web サイトを探索し、日中対訳データを収集する試みについて報告した。我々は 460 万文対の日中対訳データを収集し、翻訳精度では CCMatrix(1240 万文対) と同等であることを示した。

今後は、Common Crawl 解析から得られた対訳データ (280 万文対) から不適切な内容をフィルタリングした上で、クラウドソーシングから得られた対訳データに加えて、機械翻訳に基づく文書対応と文対応を実施し、日中対訳の品質を高める予定である。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Holger Schwenk, et al. CCMatrix: Mining billions of high-quality parallel sentences on the web. In **Proceedings of the 59th ACL**, pp. 6490–6500, 2021.

[2] Eleftheria Briakou, et al. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In **Proceedings of the 61st ACL**, pp. 9432–9452, 2023.

[3] Haoran Xu, et al. A paradigm shift in machine translation: Boosting translation performance of large language models. In **Proceedings of ICLR-2024**, 2024.

[4] Jiaxin Guo, et al. A novel paradigm boosting translation capabilities of large language models. In **Findings of the NAACL 2024**, pp. 639–649, 2024.

[5] Minato Kondo, et al. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In **Proceedings of the 21st IWSLT**, pp. 203–220, 2024.

[6] 森下睦ほか. 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. 言語処理学会第 30 回年次大会 発表論文集, 2024.

[7] Philip Resnik. Mining the web for bilingual text. In **Proceedings of the 37th ACL**, pp. 527–534. Association for Computational Linguistics, 1999.

[8] Jakob Uszkoreit, et al. Large scale parallel document mining for machine translation. In **Proceedings of the 23rd COLING**, pp. 1101–1109, 2010.

[9] Marta Bañón, et al. ParaCrawl: Web-scale acquisition of parallel corpora. In **Proceedings of the 58th ACL**, pp. 4555–4567, 2020.

[10] Makoto Morishita, et al. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proceedings of the Thirteenth LREC**, pp. 6704–6710, 2022.

[11] Ahmed El-Kishky, et al. CCAligned: A massive collection of cross-lingual web-document pairs. In **Proceedings of the 2020 EMNLP**, pp. 5960–5969, 2020.

[12] Toshiaki Nakazawa, et al. ASPEC: Asian scientific paper excerpt corpus. In **Proceedings of the Tenth LREC**, pp. 2204–2208, 2016.

[13] Holger Schwenk, et al. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In **Proceedings of the 16th EACL**, pp. 1351–1361, 2021.

[14] Chenhui Chu, et al. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese-japanese wikipedia. **ACM TALLIP**, Vol. 15, No. 2, pp. 10:1–10:22, 2015.

[15] Pierre Lison, et al. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In **Proceedings of the Eleventh LREC**, 2018.

[16] Makoto Morishita, et al. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of the Twelfth LREC**, pp. 3603–3609, 2020.

[17] Ye Kyaw Thu, et al. Introducing the Asian language treebank (ALT). In **Proceedings of the Tenth LREC**, pp. 1574–1578, 2016.

[18] Jinyi Zhang, et al. Wcc-jc 2.0: A web-crawled and manually aligned parallel corpus for japanese-chinese neural machine translation. **Electronics**, Vol. 12, No. 1140, 2023.

[19] Dániel Varga, et al. Parallel corpora for medium density languages. In **Proceedings of the RANLP-2005**, pp. 590–596, 2005.

[20] Yujie Zhang, et al. Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In **Proceedings of MTSummit X**, 2007.

[21] Víctor M. Sánchez-Cartagena, et al. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In **Proceedings of the Third WMT**, pp. 955–962, 2018.

[22] Jaume Zaragoza-Bernabeu, et al. Bicleaner AI: Bicleaner goes neural. In **Proceedings of the Thirteenth LREC**, pp. 824–831, 2022.

[23] Ruixuan Luo, et al. Pkuseg: A toolkit for multi-domain chinese word segmentation. arXiv:1906.11455, 2019.

[24] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In **Proceedings of the 16th EACL**, pp. 2112–2128, 2021.

[25] Toshiyuki Takezawa, et al. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In **Proceedings of the Third LREC**, 2002.

[26] Fangxiaoyu Feng, et al. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th ACL**, pp. 878–891, 2022.

[27] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In **Proceedings of the Eighth LREC**, pp. 2214–2218, 2012.

[28] Christian Federmann, et al. NTREX-128 – news test references for MT evaluation of 128 languages. In **Proceedings of the First Workshop on Scaling Up Multilingual Evaluation**, pp. 21–24, 2022.

[29] Tom Kocmi, et al. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In **Proceedings of the Eighth WMT**, pp. 1–42, 2023.

[30] Myle Ott, et al. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of the 2019 NAACL**, pp. 48–53, 2019.

[31] Ashish Vaswani, et al. Attention is all you need. In **Proceedings of the NeurIPS 2017**, pp. 5998–6008, 2017.

[32] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 EMNLP**, pp. 66–71, 2018.

[33] Kishore Papineni, et al. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th ACL**, pp. 311–318, 2002.

[34] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third WMT**, pp. 186–191, 2018.

[35] Ricardo Rei, et al. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 EMNLP**, pp. 2685–2702, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).