ニューラル機械翻訳のための日中対訳コーパスの拡充

張 津一 [1,2]　李 皓威 [1]　高 忠輝 [1]　毛 剣楠 [2]　田 野 [3]　松本 忠博 [2]　肖 桐 [4]
[1] 瀋陽理工大学（中国）　[2] 岐阜大学
[3] 株洲中車時代電気株式会社（中国）　[4] 東北大学（中国）
zhangjinyi@sylu.edu.cn　lihaowei21@outlook.com　gao3229@outlook.com
mao.jiannan.v9@s.gifu-u.ac.jp　tianye@csrzic.com
tad@gifu-u.ac.jp　xiaotong@mail.neu.edu.cn

概要

映画やテレビの字幕は自然言語処理のタスクにおいて重要な役割を果たしてきたが，日本語と中国語の対訳コーパスは依然として不足している．このギャップを埋めるため，著者たちは以前様々な映画やテレビシリーズのウェブサイトから字幕テキストをクロールして，日中対訳コーパス WCC-JC 1.0 とその後継の WCC-JC 2.0 を開発した．最新版の WCC-JC 3.0 は 330 万以上の対訳文対を含み，WCC-JC 2.0 から 55%以上増加した．WCC-JC 3.0 の有効性を評価するため，BLEU スコアを測定し，それに基づいてモデルの翻訳を人手で評価した．WCC-JC 3.0 は研究目的のみで利用可能である．

1　はじめに

機械翻訳は，人工知能分野の重要なタスクであり，言語障壁を克服するための最も効果的な手段の一つとされている．これは AI 技術の革新的な進展のバロメーターとして機能している．Transformer を用いたニューラル機械翻訳は，以前のフレームワークと比較して様々な言語ペアにおいて優れた翻訳結果を生み出し，現在の機械翻訳研究の主要な研究ホットスポットとして位置付けられている (1)．

高品質な日中機械翻訳出力を達成するためには，大量の日中対訳コーパスが必要である．しかし，現在の公開されている日中対訳コーパスの規模は比較的小さい．例えば，ASPEC-JC 対訳コーパスは約 68 万文対であり，他の言語ペアと比べて大きな格差がある (2)．

したがって，日中対訳コーパスを構築することによって，機械翻訳の進歩を促進する．以前の研究では，約 753K の日中文対を持つ WCC-JC 1.0 を収集した (3)．これは最初のステップであり，プロジェクトの始まりを示していた．より大きく，高品質なコーパスの必要性を認識し，次に WCC-JC 2.0 を構築した (4)．この 2.0 バージョンは，約 2,150K の文対を含むコレクションを拡張するだけでなく，慎重な人手によるアライメントを経て，全体的な品質を高めた．

WCC-JC 2.0 の進歩にも関わらず，翻訳テストと評価の結果，WCC-JC 1.0 と 2.0 はサイズと品質の面で他の利用可能なコーパスに比べて遅れていることが明らかになった．この取り組みは，WCC-JC 3.0 の拡充へと繋がった．これは中国語と日本語の話し言葉を集めた最も広範な対訳コーパスであり，世界最大の公開日中対訳コーパスとして位置付けられている．約 330 万の文対を含み，前 2.0 バージョンからの総文数が 55%の拡充を実現し，翻訳テストと人手評価でその効果を確認した．

2　関連研究

高品質の対訳コーパスは，最良の機械翻訳結果を得るための鍵である．この理解をもとに，多くの研究者がそのような対訳コーパスを作成するために努力してきた．

森下らは JParaCrawl v2.0 の規模を倍増し，21 百万以上のペアを含む JParaCrawl v3.0 をリリースした (5)．同様に，Ghaddar らと Blin らは，それぞれ英語-フランス語および日本語-フランス語の大規模なコーパスを導入し，それぞれのドメイン固有の翻訳タスクにおいて重要な役割を果たした (6; 7)．

ウェブベースのコーパスも注目を集め，Zhang らはウェブクローリング技術を利用して大規模な中英コーパスを収集し，新しいアライメント方法を強調した (8)．Jiang らは BWB コーパスでドキュメントレベルに焦点を移し，翻訳における談話の重要性を強調した (9)．

ASPEC のような先駆的な取り組みは，Nakazawa らによって科学分野に提供され，日本語-英語および中国語-日本語の広範な対訳コーパスを提供した (2)．最後に，低リソース言語に対処するため，Liu らは人手で対訳コーパス構築の方法を示し，ペルシャ語-中国語，ヒンディー語-中国語などの高品質な対訳コーパスを構築した (10)．

3 WCC-JC 3.0 の構築

3.1 ウェブクローリング

WCC-JC 1.0 及び 2.0 の方法に従い，Scrapy[1] を用いてウェブサイト[2][3] から多様な日中対訳字幕を取得し，直接ダウンロード不可の字幕は別のサイト[4] から MKV を入手し，SubtitleEdit[5] で ASS ファイルを抽出，2023 年 9 月まで更新した．歌詞は LrcHelper[6] を使用し，NetEase Cloud Music[7] から約 67,000 曲分を 12 日間でダウンロード．ニュースは Dong-a Ilbo[8] から 2000 年以降の多様なカテゴリの日中記事 100 万件以上を収集した．

3.2 対訳文の抽出

歌詞データは全曲抽出後，約 70 万文の対訳コーパスを構築した．ニュースは Dong-a Ilbo からウェブクローリングにより HTML 形式の記事を収集し，約 18 万件の日中記事を得た（約 7.5 万中国語文と 9.5 万日本語文）．図 1 に NetEase の歌詞（左）と Dong-a のニュース（右）のテキストを示す．その後，zenhan ライブラリ[9] で繁体字を簡体字に，カタカナを全角に前処理した．

3.3 テキストアライメント

歌詞テキストには元々アライメントの特性があるが，人手によるアライメント実験のために無作為に 2,000 の文対をサンプルデータとして選択した．その結果，アライメント率は 98.2%で，良いレベルであることが示された．

獲得した中国語と日本語のニュースの見出しをアライメントするために，事前学習済みのモデル LaBSE(11)[10] を利用した．そして，約 66 万文対のアライメントされたニューステキストを確認した．最終的に，字幕，歌詞，ニュースからのテキストを結合し，重複を排除して WCC-JC 3.0 を構築した．

3.4 コーパスの分割

コーパス構築の終盤で，厳選された対訳コーパスから文対をサンプリングし，開発データとテストデータに割り当てた．字幕，歌詞，ニュースの比率を考慮し，比例サンプリングを行い，テスト及び開発データ各 2,000 文対がコーパスの分布を反映するようにした．データ品質のため，10 文字以上の文対を選び，開発データとテストデータを形成し，残りは訓練用にした．図 2 には WCC-JC 3.0 の構築過程が示されており，節 3.1 から 3.4 がこれを説明している．表 1 には ASPEC-JC(2)，OpenSubtitles(12)，新 WCC-JC 3.0 の文数が記されている．

4 実験と評価

実験に使用された NMT システムについては，4.1 節で詳述している．4.2 節では，ASPEC-JC と OpenSubtitles，これら新たに導入されたコーパスを用いたそれぞれの NMT 設定で，日中翻訳の優位性を示唆する文字レベルの BLEU スコアを比較した．また，テストデータ「W」の翻訳出力を JPO 評価を使用して人手で評価した．

4.1 NMT フレームワークの設定

実験のために，fairseq をフレームワークとした (13)，fairseq から提供する transformer-iwslt-de-en という既存モデルを利用した．日本語と中国語の単語分割は，それぞれスペースがないため，MeCab[11] と Jieba[12] によって実用された．

BLEU スコアを機械翻訳出力を評価するための指標として使用された．“fairseq-score” コマンドにより，単語分割後の BLEU スコアを算出した．

4.2 評価

4.2.1 人手評価の基準

単なるアライメント分析を超えて，WCC-JC 3.0 コーパスから生じる翻訳品質を深く検証した．評価

1) https://scrapy.org/
2) https://assrt.net/
3) https://bbs.acgrip.com/
4) https://subs.kamigami.org/
5) https://github.com/SubtitleEdit
6) https://github.com/ludoux/LrcHelper
7) https://music.163.com
8) https://www.donga.com
9) https://pypi.org/project/zenhan/
10) https://github.com/UKPLab/sentence-transformers
11) http://taku910.github.io/mecab
12) http://github.com/fxsjy/jieba

<div id="lyric-content" class="bd bd-open f-brk f-ib" data-song-id="20000000" data-third-copy="false" ，copy-from>

"作词 ： 李智友"

"作曲 ： 李智友"

"宇宙探索の後での星上で"

"在太空探索之后的星球上"

"静静地星光闪烁起来"

"静かに星明りがまたたきだす"

"別れの後ですぐに信号を送った"

"在离别之后立刻发出讯号"

"地球が受信できることを願って"

"希望地球能接收到"

"在这无尽的宇宙之海"

"この果てしない宇宙の海で"

"我在寻找你的踪迹"

"君の足跡を探している"
…
…

"2035年、イ・ジウ（ソウル宇宙探査チーム）がアジア人初の宇宙飛行士として、スター・エクスプローション・エクセレンス・アワードの候補者30人に選ばれた。 スター探査優秀賞は、英語で「星の栄光」を意味し、宇宙探査の分野で世界で最も権威のある賞である。"

"世界宇宙探査機構は、「世界の宇宙探査分野で、イ・ジヨンほどミッション完遂率の高い宇宙飛行士はいない。イ・ジウは宇宙研究と惑星間探査の両分野で存在感を示している」と候補者に選んだ理由を説明している。 イ・ジウは前回の火星探査ミッションで、韓国宇宙探査チームによる初の火星着陸成功に決定的な貢献をしたが、来たる新たな宇宙探査シーズンに向けてグローバル・スペース・アライアンスに移籍した。"

"高秀研记者 xiuyan@donga.com"

"2035年，李智友（首尔太空探索队）成为首位入选全球太空探索杰出贡献奖30人候选名单的亚洲宇航员。英语中意为"星际荣耀"的全球太空探索杰出贡献奖（Star Exploration Excellence Award）是太空探索领域能获得的世界最高权威奖项。"

"全球太空探索组织表示，“在全球太空探索领域，没有一名宇航员的任务完成率比李智友更高。李智友在太空科研和星际探索中都表现出了存在感”，解释了将他选定为候选人的理由。李智友在上一次的火星探测任务中为韩国太空探索队首次成功登陆火星做出了决定性贡献，但在即将开始的新的太空探索季度，他转到了全球太空联盟。"

"高秀研記者 xiuyan@donga.com"

図 1　左：NetEase Cloud Music の歌詞テキスト；右：Dong-A Ilbo のニューステキスト（内容はダミー）.

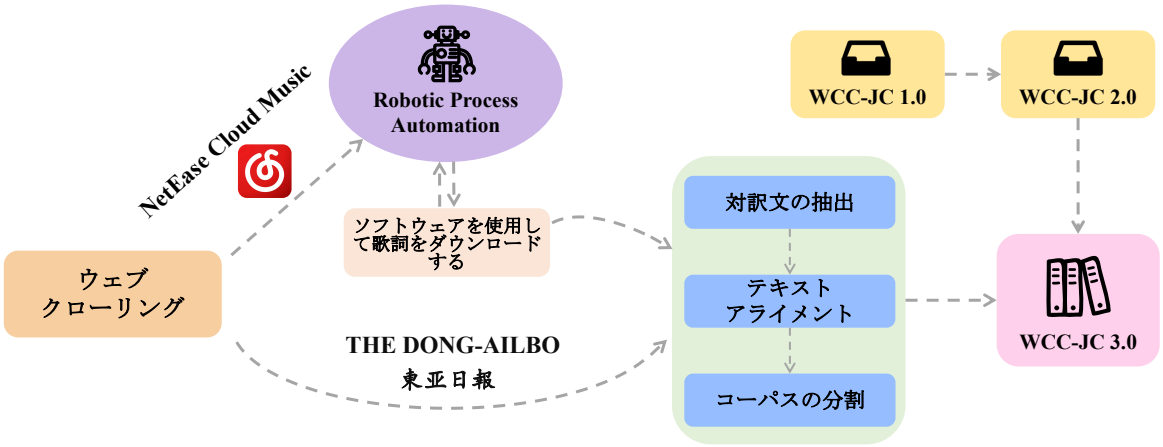

図 2　WCC-JC 3.0 のクローリングプロセス.

表 1　日中対訳コーパスのサイズの比較（カッコ内は MB）.

内容	対訳文数		
	ASPEC-JC (184.8 MB)	OpenSubtitles (72.4 MB)	WCC-JC 3.0 (381.4 MB)
訓練データ	672,315	1,087,295	3,334,828
開発データ	2,090	2,000	2,000
テストデータ	2,107	2,000	2,000

チームは，日本特許庁（JPO）[13]が設定する適切性基準を利用して，翻訳の内容忠実度を測定した．JPOは1から5までの段階を提供しており，5が最高評価である．

4.2.2 翻訳結果の分析

表 2 文字レベル日本語 → 中国語の翻訳実験結果（BLEU 値）．

訓練データ	テストデータ			
	A	O	W	N
ASPEC-JC	**34.5**	1.2	3.2	4.9
OpenSubtitles	0.0	2.1	0.0	3.0
WCC-JC 3.0	16.6	**4.1**	**20.0**	**16.9**

表 3 文字レベル中国語 → 日本語の翻訳実験結果（BLEU 値）．

訓練データ	テストデータ			
	A	O	W	N
ASPEC-JC	**44.8**	1.7	4.2	5.1
OpenSubtitles	0.1	4.0	2.3	4.1
WCC-JC 3.0	18.7	**5.1**	**26.4**	**12.8**

表 2〜3 のテストデータにおいて，ASPEC-JC は「A」，OpenSubtitles は「O」，WCC-JC 3.0 は「W」を表す．NHK「まいにち中国語」[14]の 185 文は「N」として汎用性能力の評価に使用された．翻訳モデルは fairseq 提供の transformer-iwslt-de-en を使用した．日本語 → 中国語では WCC-JC 3.0 が「A」で ASPEC-JC に及ばなかったが，他のカテゴリーでは良好な結果を示した．特に「W」と「N」では，他のコーパスと比較して顕著に高い BLEU 値を得た。中国語 → 日本語でも，WCC-JC 3.0 は「A」以外のカテゴリーでより優れた結果を示した．

4.2.3 人手評価の結果と分析

日本語 → 中国語および中国語 → 日本語の両方向の翻訳出力に対し，テストデータ「W」に注目し，人手評価を実施した．評価チームは X、Y、Z の 3 つのグループに分けられ，個々の評価結果は図 3〜4 に示されており，特定のパターンが明らかにされた．平均的に，すべてのチームが近いスコアを出したが，チーム Z は特定のカテゴリーでわずかに高いスコアを記録し，彼らの評価が他のチームに比べて少し肯定的であることを示唆した．全体的に，人手評価は賞賛に値する翻訳性能を反映し，WCC-JC 3.0 が高品質な翻訳を促進するための有用性と効率を検証した．

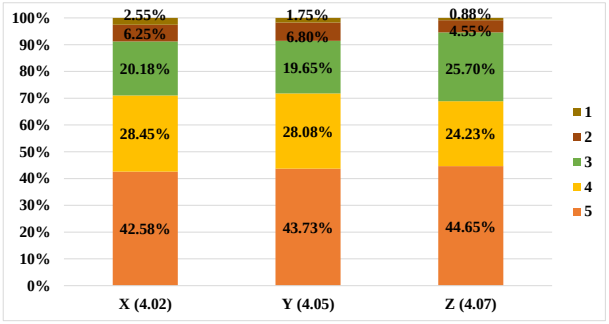

図 3 日本語 → 中国語の人手評価結果．

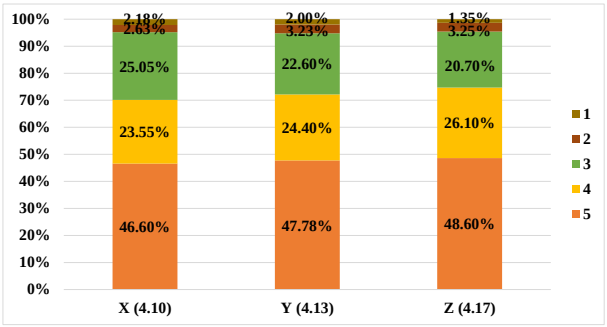

図 4 中国語 → 日本語の人手評価結果．

5 おわりに

本研究で構築した WCC-JC 3.0 は，330 万を超える日中対訳文対を有し，日中言語データベース分野における画期的な資源である．対訳字幕，歌詞，ニュースから構築されたこのコーパスは，厳格なキュレーションと精密なアライメントを経ており，話し言葉のテキストを含む広大なサイズと一般公開のアクセス性により，総括的な日中対訳コーパスとなっている．

データの著作権に関する問題に配慮し，専門家の助言を仰ぎ，WCC-JC 3.0 が中国と日本の著作権規制に合致していることを確認した．さらに，日中翻訳テストと人手評価により，コーパスの品質と多様性が証明された．将来的にはデータ拡張技術を用いてコーパスの質をさらに高め，NMT 研究を推進する．WCC-JC 3.0 は，地域経済の統一を促進し，中国と日本の相互理解を深めるための架け橋となる．

13） https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html
14） https://www.nhk.or.jp/gogaku/chinese/

謝辞

本研究にあたり，遼寧省教育庁科学研究一般若手人材プロジェクト（Grant No. LJKZ0267），瀋陽理工大学ハイレベル人材招致研究支援計画（Grant No. 1010147001004）の助成を受けた．張津一は中国国家留学基金管理委員会によって資金を提供されている（No. 202208210120）．この研究で開発されたWCC-JC 3.0 のデモ（50 万文対）は Github[15] で公開している．また，対訳コーパスの構築と評価にあたっては，多くの方々のご協力いただきました．ここに御礼申し上げます．

参考文献

[1]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**, pp. 5998–6008, 2017.

[2]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 2016, Portorož, Slovenia, May 23-28, 2016**. European Language Resources Association (ELRA), 2016.

[3]Jinyi Zhang, Ye Tian, Jiannan Mao, Mei Han, and Tadahiro Matsumoto. Wcc-jc: A web-crawled corpus for japanese-chinese neural machine translation. **Applied Sciences**, Vol. 12, No. 12, 2022.

[4]Jinyi Zhang, Ye Tian, Jiannan Mao, Mei Han, Feng Wen, Cong Guo, Zhonghui Gao, and Tadahiro Matsumoto. Wcc-jc 2.0: A web-crawled and manually aligned parallel corpus for japanese-chinese neural machine translation. **Electronics**, Vol. 12, No. 5, 2023.

[5]Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. Jparacrawl v3.0: A large-scale english-japanese parallel corpus. **arXiv preprint arXiv:2202.12607**, 2022.

[6]Abbas Ghaddar and Philippe Langlais. SEDAR: a large scale french-english financial domain parallel corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020**, pp. 3595–3602. European Language Resources Association, 2020.

[7]Raoul Blin and Fabien Cromières. Cjafr-v3 : A freely available filtered japanese-french aligned corpus. **CoRR**, Vol. abs/2208.13170, , 2022.

[8]Yunhui Zhang. Key technologies of constructing parallel corpus for chinese-english intertranslation based on web. In **2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)**, pp. 940–944. IEEE, 2023.

[9]Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. A bilingual parallel corpus with discourse annotations. **CoRR**, Vol. abs/2210.14667, , 2022.

[10]Yan Liu and Deyi Xiong. Construction method of parallel corpus for minority language machine translation. **Computer Science**, Vol. 49, No. 1, pp. 41–46, 2022.

[11]Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**, pp. 878–891. Association for Computational Linguistics, 2022.

[12]Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018**. European Language Resources Association (ELRA), 2018.

[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 **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.

15） https://github.com/zhang-jinyi/Web-Crawled-Corpus-for-Japanese-Chinese-NMT