ニューラル機械翻訳のための日中対訳コーパスの構築

張 津一 [1,2]　田 野 [3]　韓 梅 [4]　毛 剣楠 [2]　松本 忠博 [2]
[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] 湖南工業大学 （中国）
zhangjinyi@sylu.edu.cn　tianye@csrzic.com
hanmei@hut.edu.cn　z4525087@edu.gifu-u.ac.jp　tad@gifu-u.ac.jp

概要

現在，ニューラル機械翻訳の学習データとして利用可能な，ある程度の規模を持つ日本語-中国語対訳コーパスは少ない．とくに日常会話などの話し言葉を主な対象としたものは限られている．本研究では Web 上に存在する TV 番組などの字幕データをクローリングにより収集して，一定の規模の日中対訳コーパスの構築を試みた．このコーパス WCC-JC（Web Crawled Corpus - Japanese and Chinese）で学習した翻訳モデルを使って翻訳した結果を人手で評価し，品質・効果を確認した．

1　はじめに

機械翻訳は人工知能の重要な分野であり，原言語を出力言語に翻訳する方法を研究することは，言語の壁の問題を解決することができる最も効果的な手段の一つである．長年の開発を経て，ニューラル機械翻訳（Neural Machine Translation，NMT）は様々な言語ペアで従来のフレームワークよりも優れた翻訳結果を出しており，大きな可能性を秘めた新しい機械翻訳モデルとなっている．NMT では，学習データの量が翻訳結果の質を大きく左右するものの，大規模な翻訳モデルを学習することができる．

機械翻訳の分野では，日本語と中国語という 2 つの言語が複雑に絡み合っているため，日中の機械翻訳は難しい．高品質な翻訳を得るためには，日中翻訳コーパスのデータ量が大量に必要となる．しかし，英語を含む言語対や欧州の言語間の言語対などが数千万から数億文対である．そして，様々な分野の文が含まれる．これに対し，公開された日中対訳コーパスは多くない．例えば，JPO 日中対訳コーパスは 1.3 億件（約 26GB）および 0.1 億件（約 1.4GB）があるが，データの全てを用いた研究成果の提出が必要である[1)]．このコーパスの内容の種類は特許のみである．ASPEC-JC の収録件数は，約 67 万文である [1]．このコーパスの内容の種類は科学論文の概要のみである．上記の 2 つは，公開されているコーパスの中でも非常に規模の大きなコーパスである．だが，日常的な文の翻訳に適した日中対訳コーパスは，まだそれほど多くない．

現在の研究のほとんどは，書き言葉のカテゴリーに入る文の翻訳を対象にしている．一方，話し言葉の場合，省略などにより書き言葉よりも曖昧さが多くなり，文脈を把握する必要性も高まる．文の長さも書き言葉より短いのが普通である．話し言葉でも，俗語や方言があるものは，従来の翻訳者では見落とされることがある．また，音声の入力から翻訳結果の出力までを行うマルチモーダル翻訳が近い将来のトレンドになると言われるが，その実現には話し言葉の翻訳に適した対訳コーパスが必要である．

本研究では，Web をクロールして作成した日中対訳コーパスとその評価について述べる．作成したコーパス WCC-JC は，インターネット上からクロールされた映画と TV 番組の字幕データの日中両言語の対応をとることで構築されている．既存の日中対訳コーパスではあまり扱われてこなかった話し言葉の対訳も対象している．

2　関連研究

現在，公開されている日中対訳翻訳コーパスは非常に少なく，JPO 日中対訳コーパス，ASPEC-JC[1] と TED Talks[2)3)] に含まれる日中対訳コーパスなどがあるのみである．

Lavecchia らは，映画の字幕ファイルから対訳コーパスを自動的に構築する方法を提案し，それを用い

1）https://alaginrc.nict.go.jp/jpo-outline.html
2）https://ted.com
3）https://github.com/ajinkyakulkarni14/TED-Multilingual-Parallel-Corpus

て対訳辞書も作成している [2].

Chu らは，Wikipedia から 126k 以上の日中対訳コーパスを構築するための対訳文抽出システムを提案した [3].

Pryzant らは，インターネット上からクロールされた映画と TV 番組の字幕データを日英対応させることで構築された．JESC は，自由に利用できる日英対訳コーパスの中で最大規模（約 280 万文）のコーパスである [4].

OpenSubtitles2018 は映画字幕データの多言語対訳コーパスである [5]．日英対訳コーパスはおよそ 2,000 件の映画からなる 200 万文の対訳コーパスであり，機械翻訳の分野や，映画字幕という特徴を活かしたタスクにおける利用が検討される．

森下らは，大規模 Web ベース 1000 万文対以上の日英対訳コーパス JParaCrawl を紹介した．無料で一般に公開した [6].

Guokun らはインターネットから言語資源をクロールしてコーパスを自動的に構築したが，データのフィルタリングがうまくいっていなかった [7].

中澤らは，「ビジネス」における「対話」を対訳コーパスのドメインとして構築した BSD コーパスおよび WAT2020 での BSD 翻訳タスクの結果の紹介を行い，翻訳結果の分析から対話翻訳の課題について議論した [8].

また，日中翻訳では，Zhang らが対訳コーパスの資源不足を考慮し，日中翻訳の品質を向上させるコーパス拡張アプローチを提案した [9].

以上の関連研究により，コーパスは翻訳精度の向上と言語処理の他の分野に重要な役割を果たす．このように，ニューラル機械翻訳のためのオープンな日中対訳コーパスの構築は資源不足問題に大きな意味を持っている．本研究では，日中機械翻訳を発展させ，ある程度の規模を持つ日中対訳コーパスの作成を目標とする．

3 日中対訳コーパスの構築

3.1 対訳文抽出

多くの日中対訳文を含む Web サイトを選び，映画・ドラマと TV 番組の字幕ファイルを含む Web サイトから Scrapy[4] を利用し，字幕ファイルを取得する．字幕ファイルの中に，俗語，話し言葉，説明文，物語解説の対訳がある．これらは既存のコーパスではあまり扱われてこなかった分野である．

取得した字幕ファイルのほとんどは ASS 形式のファイルである．図 1 に ASS ファイルの内容の一例を表す．図のように ASS ファイルでは Dialogue 行に，字幕の表示内容と表示時間，表示スタイルなどの情報が記載されている．言語情報は字幕のレイヤー情報（図中では先頭の“0”）や表示スタイル（図中では“DefaultJp”）に現れる場合が多い．本研究では，字幕表示スタイルの文字列から日本語か中国語かを推定し，字幕表示開始時間と字幕表示終了時間をもとに日中のセリフの対応関係を判断した．

データ収集後，中国語文中の繁体字を簡体字に，日本語文中の半角カタカナを全角カタカナに統一し，重複する文対を除去した．

3.2 コーパスの分割

収集した文対のうち，10 文字以上の文対をテストデータと検証データ（開発データ）用にそれぞれ 2000 文対ランダムに抽出し，残りを訓練データとした．表 1 は構築されたコーパスと ASPEC-JC，OpenSubtitles と WCC-JC コーパスの文対数を表す．データサイズ（バイト数）は ASPEC-JC > OpenSubtitles > WCC-JC の順だが，文数は ASPEC-JC よりも多いことが分かる．

4 実験と評価

本コーパスの有効性を確認するために翻訳実験を行った．以下，実験に使用した NMT システムの設定（4.1 節），ASPEC-JC，OpenSubtitles，本研究のコーパスのデータセットで学習した NMT モデルでの翻訳精度（BLEU スコア）（4.2 節），JPO スコアを用いて，テストデータの翻訳結果の人手での評価（4.3 節）について述べる．日中翻訳では，文字レベルと LSTM モデルが最も効果的であると言われている [10] が，本実験では文字レベルと subword レベル，LSTM モデルと Transformer モデルを使用した．

4.1 NMT システムの設定

本実験では，fairseq[11] を用いて NMT モデルを訓練する．fairseq の 2 つの事前設定されたアーキテクチャ，lstm-wiseman-iwslt-de-en（表の中，LSTM と略す）と transformer-iwslt-de-en (表の中，Transformer と略す) を使用する．Subword レベルの実現には subword-nmt[5] を使用する．

4） https://scrapy.org/

5） https://github.com/rsennrich/subword-nmt

```
Dialogue: 0,1:44:02.47,1:44:05.56,DefaultJp,,0,0,0,,{\blur4}すみません。タクシー一台急いでお願いします。
Dialogue: 0,1:44:08.61,1:44:13.53,DefaultJp,,0,0,0,,{\blur4}すぐ行きますが、何か目印を教えてください。
Dialogue: 0,1:44:14.19,1:44:16.66,DefaultJp,,0,0,0,,{\blur4}山手通りを渋谷方面へ来てください。
Dialogue: 0,1:44:48.27,1:44:50.40,DefaultJp,,0,0,0,,{\blur4}そうすると、东急线の踏切があるから、それを渡ると、大きな交差点に出ます。
Dialogue: 0,1:44:50.98,1:44:54.99,DefaultJp,,0,0,0,,{\blur4}路线の下を通るんですか。
Dialogue: 0,1:44:56.28,1:45:01.58,DefaultJp,,0,0,0,,{\blur4}じゃなくて、踏切。
Dialogue: 0,1:45:02.45,1:45:03.16,DefaultJp,,0,0,0,,{\blur4}で、初めの大きな交差点を右ね。
Dialogue: 0,1:45:09.18,1:45:10.06,DefaultJp,,0,0,0,,{\blur4}しばらく来ると、左側にお寺があります。
Dialogue: 0,1:45:17.69,1:45:20.57,DefaultJp,,0,0,0,,{\blur4}その手の前の细い道を左。
Dialogue: 0,1:45:23.53,1:45:23.78,DefaultJp,,0,0,0,,{\blur4}突きあたったら、右に曲がってください。
Dialogue: 0,1:45:24.82,1:45:25.45,DefaultJp,,0,0,0,,{\blur4}三轩目の左側の家です。
Format: Layer, Start,    End,    Style, Name, MarginL, MarginR, MarginV, Effect, Text
```

図 1　ASS ファイル中身の一例（内容はダミー）

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 (184.8MB)	OpenSubtitles (72.4MB)	WCC-JC (52.2MB)
訓練データ	672,315	1,087,295	749,017
開発データ	2,090	2,000	2,000
テストデータ	2,107	2,000	2,000

4.2　翻訳結果

表 2〜表 5 において，A は ASPEC-JC のテストデータ，B は OpenSubtitles のテストデータ，C は WCC-JC のテストデータ，D は翻訳モデルの汎化能力を検証するために，NHK ラジオ『まいにち中国語』[6] のテキストから抽出した 185 文，テストデータとして表す．

表 2〜表 5 を見ると，訓練データとテストデータの組合せではそれぞれ自身のテストデータで最も高い BLEU 値を取得したことがわかる．J→C 文字レベルの Transformer モデルでは，我々の WCC-JC は OpenSubtitles 自身による翻訳よりも効果的であり，WCC-JC の汎化性が示された．また，C，D のいずれにおいても，WCC-JC は ASPEC-JC や OpenSubtitles よりも良い結果を得ることができた．さらに，各表の結果を比較すると，基本的には J→C、C→J のいずれの方向においても，文字レベルの Transformer モデルがより有効であることがわかる．

4.3　人手による評価

人手評価基準として，特許庁が公開している「特許文献機械翻訳の品質評価手順」[7] の中の「内容の伝達レベルの評価」（JPO スコア）を採用した．これは機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを，参照訳の内容に照らして 5 段階（評価値 5 が最もよく，1 が最も悪い）の評価基準で主観的に評価するものである．

WCC-JC のテストデータ C については，J→C と C→J の文字レベル Transformer の 2 つの実験で最も高い BLEU 値が観測された．より良い評価を行うために，JPO スコアを用いてこの 2 つの翻訳結果を人手評価することにした．評価は A，B，C の 3 つの評価者グループで分担した（評価者たちは日本留学の経験があり，日本語能力試験 N2 以上の資格を持つ）．各グループは 2 人ずつ独立に評価し，平均を取った．図 2 と図 3 は J→C、C→J の人手評価結果（グラフ内の数字は各評価値の割合を，グループ名の後の括弧内の数字は評価値の平均を表す）．JPO スコアの平均値について，J→C では 3.87，3.90，4.00，C→J で 3.97，4.04，4.07，高い結果になった．WCC-JC コーパスの性質上，非常に短い文が多いの

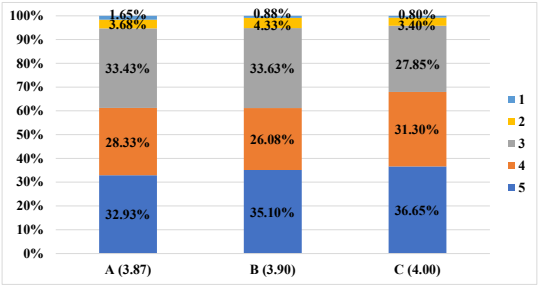

図 2　人手による評価の結果（日本語 → 中国語）

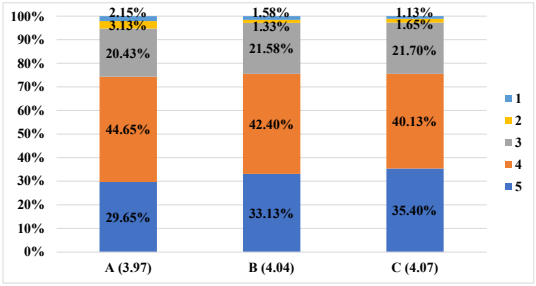

図 3　人手による評価の結果（中国語 → 日本語）

6）　https://www2.nhk.or.jp/gogaku/onayami/chinese/
7）　https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html

表 2 文字レベル日本語 → 中国語翻訳実験結果（BLEU 値）

Data\Method	Character Level							
	LSTM				Transformer			
Test data →	A	B	C	D	A	B	C	D
ASPEC-JC	**32.0**	1.0	2.6	1.9	**34.5**	1.2	3.2	4.9
OpenSubtitles	0.3	**5.0**	2.4	7.1	0.0	2.1	0.0	3.0
WCC-JC	2.6	2.7	**12.1**	**9.5**	3.2	**3.9**	**15.9**	**13.7**

表 3 Subword レベル日本語 → 中国語翻訳実験結果（BLEU 値）

Data\Method	Subword Level							
	LSTM				Transformer			
Test data →	A	B	C	D	A	B	C	D
ASPEC-JC	**31.3**	1.1	2.3	3.0	**34.0**	1.2	3.3	6.6
OpenSubtitles	0.1	**4.3**	2.1	5.7	0.0	**3.5**	0.9	4.8
WCC-JC	1.9	2.4	**11.3**	**7.1**	2.0	2.9	**14.3**	**10.3**

表 4 文字レベル中国語 → 日本語翻訳実験結果（BLEU 値）

Data\Method	Character Level							
	LSTM				Transformer			
Test data →	A	B	C	D	A	B	C	D
ASPEC-JC	**38.8**	1.2	2.8	2.8	**44.8**	1.7	4.2	5.1
OpenSubtitles	0.2	**5.5**	4.0	5.6	0.1	**4.0**	2.3	4.1
WCC-JC	3.2	3.4	**13.8**	**8.5**	3.5	3.9	**17.1**	**9.0**

表 5 Subword レベル中国語 → 日本語翻訳実験結果（BLEU 値）

Data\Method	Subword Level							
	LSTM				Transformer			
Test data →	A	B	C	D	A	B	C	D
ASPEC-JC	**39.8**	1.1	2.9	3.3	**44.3**	1.4	4.0	5.3
OpenSubtitles	0.2	**4.2**	2.6	3.6	0.2	**4.6**	2.6	4.0
WCC-JC	2.3	2.7	**13.0**	**6.4**	2.8	2.3	**15.3**	**7.7**

で，この結果がどの程度満足のいくものなのかはより深く分析する必要がある．また，別の実験についても評価を行い，結果の違いとその理由を明らかにする必要がある．

5 おわりに

本研究では，日中対訳コーパス WCC-JC を紹介した．本コーパスは Web をクロールし，日中対訳文を自動的に収集し，作成した．最終的に約 75 万文対の日中対訳データが得られた．本コーパスは現時点で一般に利用できる日中コーパスの中では規模が大きく，既存のコーパスではあまり扱われてこなかった話し言葉の対訳も対象している．

語学講座のテキストから抽出した会話文を用いた実験では，BLEU 値は低いものの，比較した日中コーパスの中では最も高い精度で翻訳できることを確認した．人手による評価も高いものであった．

今後の課題として，より大規模な Web クロール，コーパス作成を行うことが挙げられる．さらに，対訳文のアライメントを高精度化することも重要な課題である．また，今後より多くの言語対に対応していくことも検討している．

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（No.LJKZ0267），瀋陽理工大学ハイレベル人材招致研究支援計画（No.1010147001004）の助成を受けています．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] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC 2016)**, pp. 2204–2208, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

[2] Caroline Lavecchia, Kamel Smaili, and David Langlois. Building a bilingual dictionary from movie subtitles based on inter-lingual triggers. In **Proceedings of Translating and the Computer 29**, London, UK, November 29-30 2007. Aslib.

[3] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Constructing a Chinese—Japanese parallel corpus from Wikipedia. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 642–647, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[4] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English Subtitle Corpus. **Language Resources and Evaluation Conference (LREC)**, 2018.

[5] 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 2018. European Language Resources Association (ELRA).

[6] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[7] Guokun Lai, Zihang Dai, and Yiming Yang. Unsupervised parallel corpus mining on web data. **ArXiv**, Vol. abs/2009.08595, , 2020.

[8] 中澤敏明, 李凌寒, MatıssRikters. ビジネスシーン対話対訳コーパスの構築と対話翻訳の課題. 第 27 回年次大会 発表論文集, pp. 1375–1380. 言語処理学会, 2021.

[9] Jinyi Zhang and Tadahiro Matsumoto. Corpus augmentation for neural machine translation with chinese-japanese parallel corpora. **Applied Sciences**, Vol. 9, No. 10, 2019.

[10] Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. Is word segmentation necessary for deep learning of Chinese representations? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3242–3252, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] 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 NAACL-HLT 2019: Demonstrations**, 2019.