JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus

Makoto Morishita[1], Katsuki Chousa[1], Jun Suzuki[2], Masaaki Nagata[1]

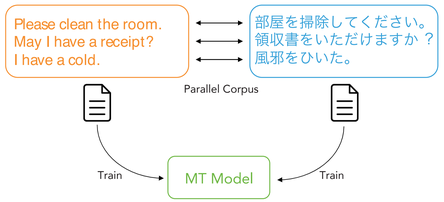
[1]NTT Communication Science Laboratories, NTT Corporation, [2]Tohoku University

Abstract

Current MT models are mainly trained with parallel corpus.
→ Corpus quality and quantity are important for accuracy.
→ However, current available English-Japanese parallel corpora are limited.
It is one of the severe issues for this language pair.

Contributions

We created a large parallel corpus for English-Japanese from the web.
→ Our corpus exceeds 21M sentences, twice as large as the previous JParaCrawl v2.0.
→ It is now publicly available on our website.

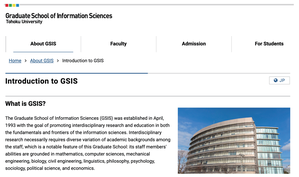


JParaCrawl v3.0

How to create

1. Find websites written in parallel based on the CommonCrawl.
   → List 100K domains where the En and Ja data sizes are similar.
2. Crawl listed websites.
3. Extract the parallel sentences from the crawled data
   → We used the machine translation-based aligning method "bleualign."
4. Filter the noisy sentence pairs based on the rules, dictionary, and language models.

https://www.is.tohoku.ac.jp/jp/introduction/outline.html



https://www.is.tohoku.ac.jp/en/introduction/outline.html

Results

Combined with JParaCrawl v2.0, our corpus exceeds 21 million sentences.
→ We released it as JParaCrawl v3.0.

Version	# Sentences	Creation date
v1.0	4,817,172	Nov. 2019
v2.0	8,809,771	Jan. 2020
v3.0	21,891,738	Dec. 2021

Experiments

BLEU scores

		English-Japanese				Japanese-English			
Test set	**Domain**	**v1.0**	**v2.0**	**v3.0**	**v3.0-v2.0**	**v1.0**	**v2.0**	**v3.0**	**v3.0-v2.0**
ASPEC	Scientific paper	24.7	26.5	**27.0**	+0.5	18.3	19.7	**21.0**	+1.3
JESC	Movie subtitle	6.6	6.5	**6.8**	+0.3	7.0	7.5	**8.3**	+0.8
KFTT	Wikipedia article	17.1	**18.9**	17.9	-1.0	13.7	16.2	**17.0**	+0.8
TED	TED talk	11.5	12.6	**13.0**	+0.4	11.0	11.9	**12.0**	+0.1
Business Scene Dialogue corpus	Dialogue	12.4	13.5	**14.1**	+0.6	17.4	19.6	**19.9**	+0.3
WMT20 News En-Ja	News	20.7	21.9	**23.5**	+1.6	21.3	23.3	**24.0**	+0.7
WMT20 News Ja-En	News	20.1	22.8	**23.7**	+0.9	19.2	21.0	**21.6**	+0.6
WMT21 News En-Ja	News	21.1	21.8	**25.1**	+3.3	21.9	23.1	**23.9**	+0.8
WMT21 News Ja-En	News	19.6	21.5	**22.8**	+1.3	18.1	20.7	**21.7**	+1.0
WMT19 Robustness En-Ja	SNS	12.4	12.5	**14.4**	+1.9	15.6	16.8	**17.2**	+0.4
WMT19 Robustness Ja-En	SNS	11.5	12.3	**13.0**	+0.7	16.0	17.2	**17.7**	+0.5
WMT20 Robustness Set1 En-Ja	Wikipedia comments	15.2	15.8	**18.7**	+2.9	20.0	20.6	**21.4**	+0.8
WMT20 Robustness Set2 En-Ja	SNS	12.7	13.0	**14.5**	+1.5	16.4	**17.4**	17.2	-0.2
WMT20 Robustness Set2 Ja-En	SNS	7.9	8.2	**8.9**	+0.7	12.0	12.6	**13.8**	+1.2
IWSLT21 Simultaneous Translation En-Ja	TED talk	12.5	13.3	**14.5**	+1.2	12.9	14.3	**15.0**	+0.7

Question: Does the new JParaCrawl v3.0 boost the translation accuracy?
Result: Yes. We confirmed the gain on various test sets.

Translation Example

Source	院内に「濃厚接触者」はいませんが、接触者全員に PCR 検査を実施し、 女性が関係した病棟などを閉鎖して徹底的に消毒するということです。
Reference	There are no known “close contacts” in the hospital, but all contacts will be subjected to PCR tests, and the wards and other areas where the women had been will be closed and thoroughly disinfected.
JParaCrawl v1.0	There is no “strong contact person” in the hospital, but a PCR test will be conducted for all the contacts, and women will close the wards and thoroughly disinfect them.
JParaCrawl v2.0	Although there is no “strong contact person” in the hospital, PCR tests will be performed on all contact persons, and the wards related to women will be closed and thoroughly disinfected.
JParaCrawl v3.0	There are no “close contacts” in the hospital, but PCR tests will be conducted for all contacts, and the wards related to women will be closed and thoroughly disinfected.

JParaCrawl v3.0 is based on the latest web.
Thus the new model can correctly translate the newly used term “close contacts.”