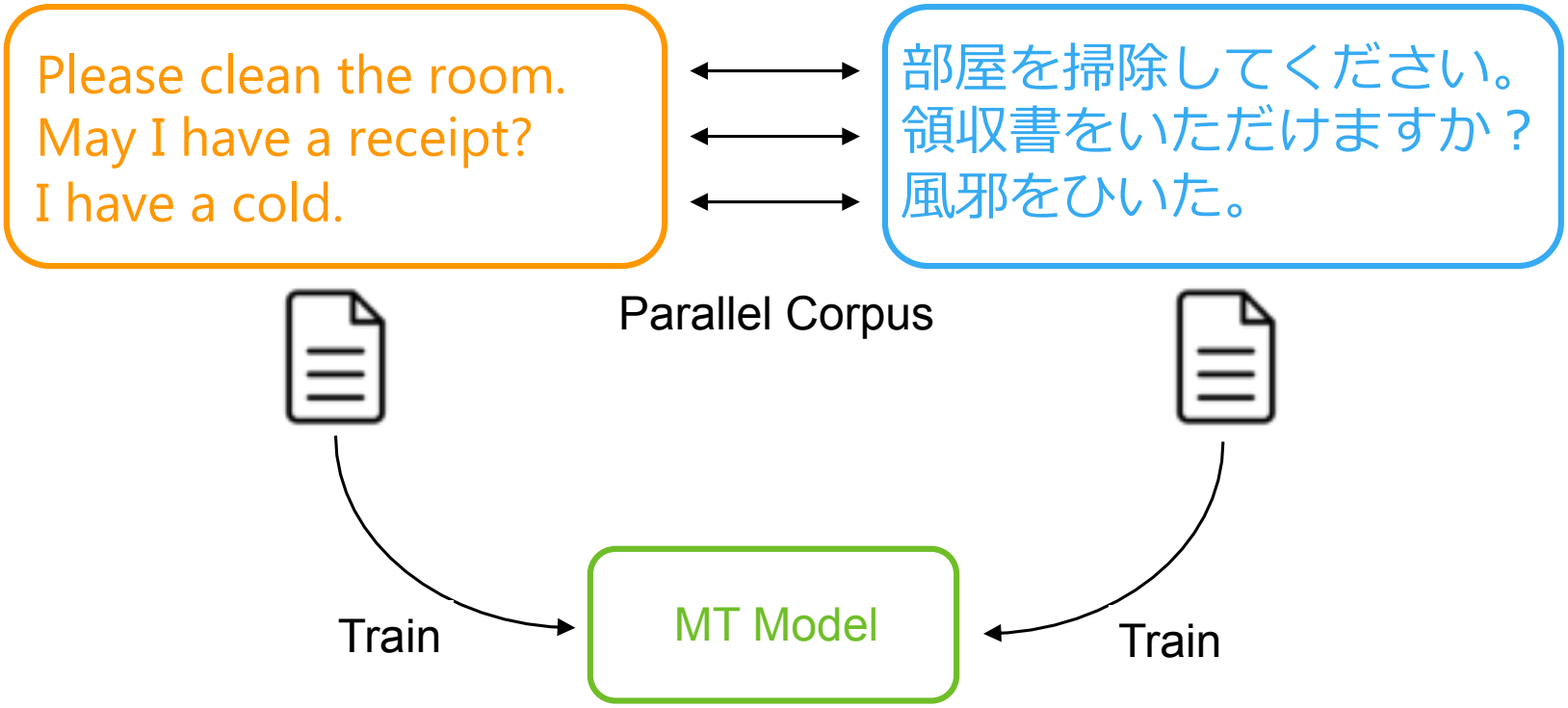
JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata

NTT Communication Science Laboratories

LREC 2022

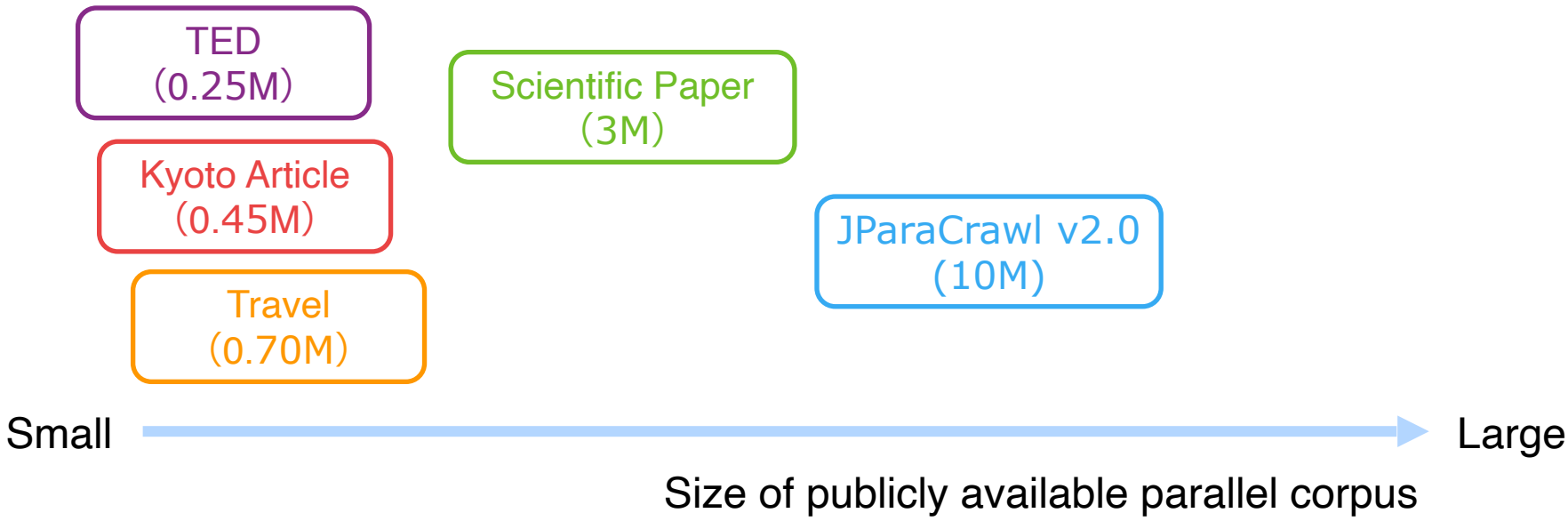
Training of Machine Translation Models



- Current MT models are mainly trained with parallel corpus
  - Corpus quality and quantity are important for accuracy

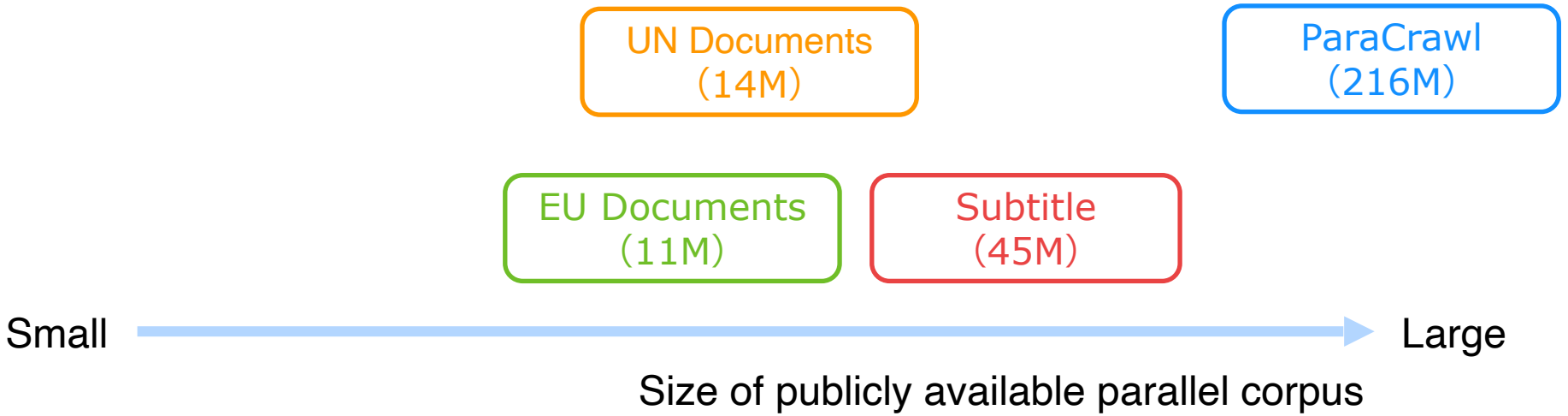
Current English-Japanese Parallel Corpora

- Current available parallel corpora are still limited



Current English-French Parallel Corpora

- Compared to the En-Ja, there are a lot of parallel corpora available.



Our Purpose

- We want to boost the En-Ja/Ja-En translation accuracy.
- We want to make En-Ja more major language pair.

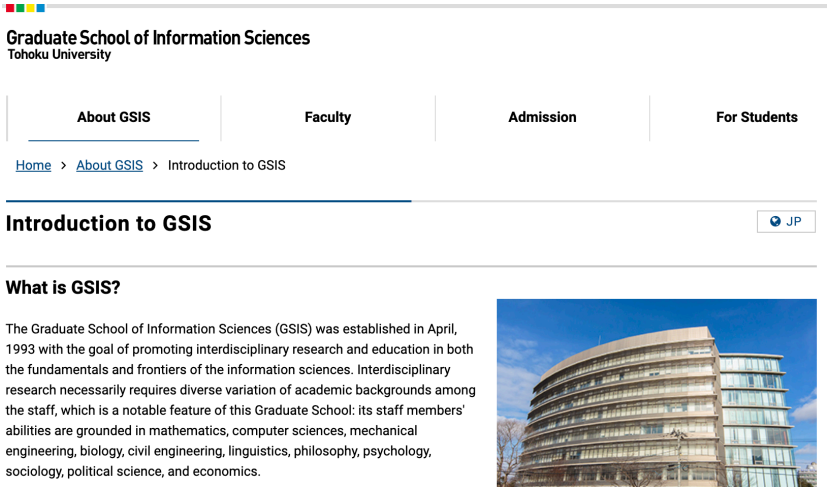
Thus we created a large-scale parallel corpus: JParaCrawl.

How we crawl

Crawling Parallel Sentences from the Web

- Find the websites that have parallel En/Ja sentences

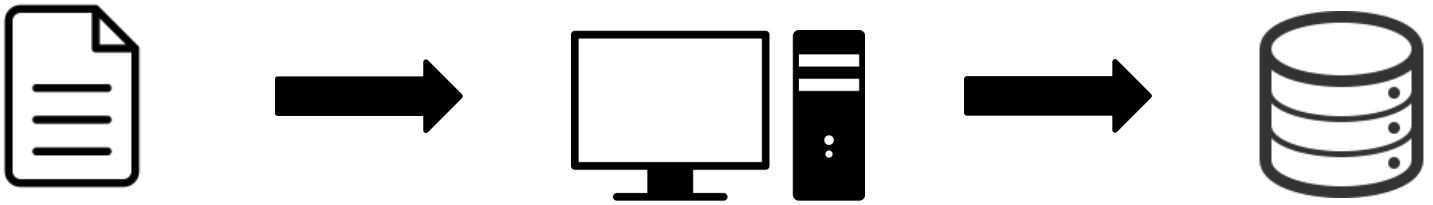
https://www.is.tohoku.ac.jp/jp/introduction/outline.html



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

Pipeline

1. **Detect languages on CommonCrawl**
   - Make language stats to find which website has bilingual texts



CommonCrawl

Language detection with CLD2

Language stats

Domain	English [KB]	Japanese [KB]
xxx.jp	1000	2000
yyy.com	45000	100
zzz.co.jp	3000	2500
…	…	…

Pipeline

2. List crawling candidate domains

- We focused on the ratio between English and Japanese
- listed 100,000 candidate domains, which roughly have the same ratio.

3. Crawl the candidate websites.

Pipeline

4. Extract parallel sentences from the crawled data

- Find sentence pairs with the MT system.

5. Filter the noisy sentences

- Based on the heuristics, lexicons, language models

Results

- Combined with JParaCrawl v2.0,
  our corpus exceeds 21 million sentences.
  - This is twice as large as the previous JParaCrawl v2.0.
- We released it as JParaCrawl v3.0.
  - It is now publicly available on our website.
    - http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

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

Experiments

- Trained the model and tested it on the various domains

  → Check which domain our new data could boost the most.

Test set	Domain
ASPEC	Scientific paper
JESC	Movie subtitle
KFTT	Wikipedia article
TED	TED talk
Business Scene Dialogue corpus	Dialogue
WMT20 News En-Ja	News
WMT20 News Ja-En	News
WMT21 News En-Ja	News
WMT21 News Ja-En	News
WMT19 Robustness En-Ja	SNS
WMT19 Robustness Ja-En	SNS
WMT20 Robustness Set1 En-Ja	Wikipedia comments
WMT20 Robustness Set2 En-Ja	SNS
WMT20 Robustness Set2 Ja-En	SNS
IWSLT21 Simultaneous Translation En-Ja	TED talk

Translation Results (En-Ja)

Test set	Domain	v1.0	v2.0	v3.0	v3.0-v2.0
ASPEC	Scientific paper	24.7	26.5	**27.0**	+0.5
JESC	Movie subtitle	**6.6**	6.5	6.8	+0.3
KFTT	Wikipedia article	17.1	**18.9**	17.9	-1.0
TED	TED talk	11.5	12.6	**13.0**	+0.4
Business Scene Dialogue corpus	Dialogue	12.4	13.5	**14.1**	+0.6
WMT20 News En-Ja	News	20.7	21.9	**23.5**	+1.6
WMT20 News Ja-En	News	20.1	22.8	**23.7**	+0.9
WMT21 News En-Ja	News	21.1	21.8	**25.1**	+3.3
WMT21 News Ja-En	News	19.6	21.5	**22.8**	+1.3
WMT19 Robustness En-Ja	SNS	12.4	12.5	**14.4**	+1.9
WMT19 Robustness Ja-En	SNS	11.5	12.3	**13.0**	+0.7
WMT20 Robustness Set1 En-Ja	Wikipedia comments	15.2	15.8	**18.7**	+2.9
WMT20 Robustness Set2 En-Ja	SNS	12.7	13.0	**14.5**	+1.5
WMT20 Robustness Set2 Ja-En	SNS	7.9	8.2	**8.9**	+0.7
IWSLT21 Simultaneous Translation En-Ja	TED talk	12.5	13.3	**14.5**	+1.2

We could see the v3.0 model surpassed the previous model on almost all the test sets.

Translation Results (En-Ja)

Test set	Domain	v1.0	v2.0	v3.0	v3.0-v2.0
ASPEC	Scientific paper	24.7	26.5	**27.0**	+0.5
JESC	Movie subtitle	**6.6**	6.5	6.8	+0.3
KFTT	Wikipedia article	17.1	**18.9**	17.9	-1.0
TED	TED talk	11.5	12.6	**13.0**	+0.4
Business Scene Dialogue corpus	Dialogue	12.4	13.5	**14.1**	+0.6
WMT20 News En-Ja	News	20.7	21.9	**23.5**	+1.6
WMT20 News Ja-En	News	20.1	22.8	**23.7**	+0.9
WMT21 News En-Ja	News	21.1	21.8	**25.1**	+3.3
WMT21 News Ja-En	News	19.6	21.5	**22.8**	+1.3
WMT19 Robustness En-Ja	SNS	12.4	12.5	**14.4**	+1.9
WMT19 Robustness Ja-En	SNS	11.5	12.3	**13.0**	+0.7
WMT20 Robustness Set1 En-Ja	Wikipedia comments	15.2	15.8	**18.7**	+2.9
WMT20 Robustness Set2 En-Ja	SNS	12.7	13.0	**14.5**	+1.5
WMT20 Robustness Set2 Ja-En	SNS	7.9	8.2	**8.9**	+0.7
IWSLT21 Simultaneous Translation En-Ja	TED talk	12.5	13.3	**14.5**	+1.2

We could see the v3.0 model surpassed the previous model on almost all the test sets.

→ Especially works well in the news domain

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.”

Conclusion

- We created a large-scale English-Japanese parallel corpus called JParaCrawl v3.0.
- It contains more than 21 million sentence pairs.
  - It is now available on our website.
- From the experiments, it boosts the translation accuracy in the various domains, especially in the news domain.

END