07

Large-scale parallel corpus construction technology

# Look for bilingual sentence pairs in the world

## Abstract

Machine translation systems require many bilingual sentence pairs (translations of each other) as training data. We are researching technology to build a parallel corpus (bilingual database) by collecting bilingual text scattered on the Internet (Web) and in patent application archives. JParaCrawl, a web-based parallel corpus, was constructed by efficiently collecting many bilingual sentence pairs from the Web using crowdsourcing. JaParaPat, a patent parallel corpus, has improved the quality of its sentence pairs by alternately extracting data and training models. Both are the world's largest publicly available parallel corpus between Japanese and English. We will further enhance our technology to automatically build a high-quality parallel corpus in specific fields, such as medicine and finance, which are rich in specialized terminology, and in particular language pairs, such as Chinese and Japanese, to implement a machine translation system customized to the needs of our customers.

## Machine translation and parallel data

The quantity and quality of parallel data determines machine translation accuracy

言語はコミュニケーションの手段である

Translator → Translation model → Translation model training ← Parallel corpus

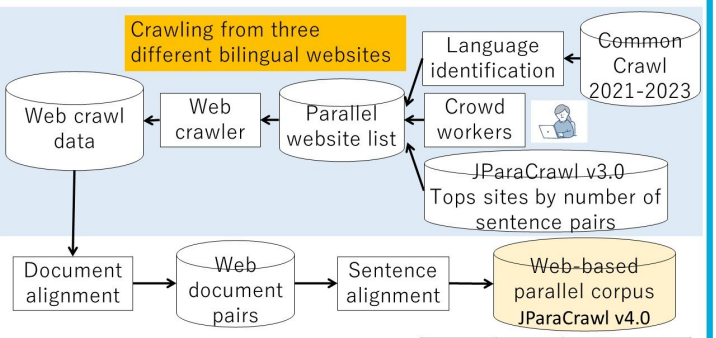
首相は同省に徹底調査を言明した
The prime minister strictly ...

日本を取り巻く安全環境は急速に…
The security environment ...

Language is a means of communication

Building a large and high-quality parallel corpus from the Internet (Web) and patents

## Web-based parallel corpus: JParaCrawl v4.0

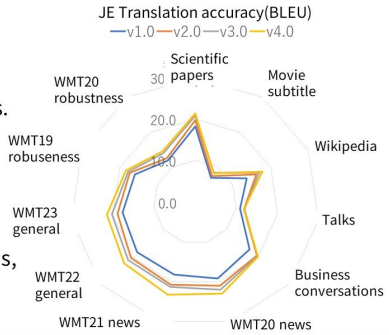



| Website mining | #sites | parallel | #sents |
|---|---|---|---|
| CommonCrawl | 50K | 35% | 22.8M |
| Crowd workers | 21K | 85% | 16.9M |
| v3.0 top sites | 24K | 90% | 6.9M |

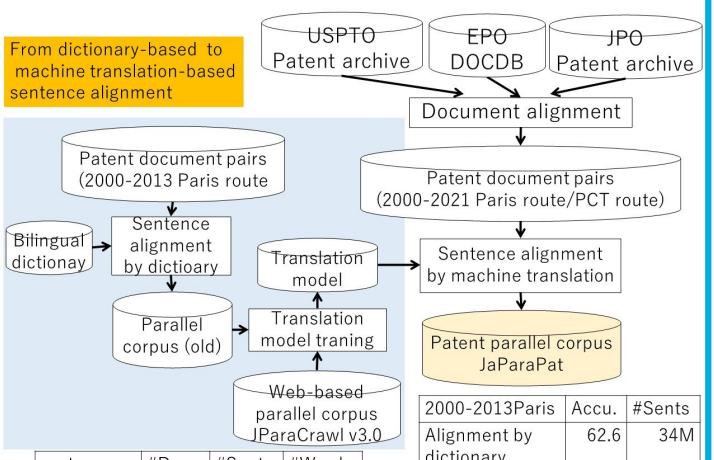
| version | #sents | Creation time |
|---|---|---|
| v1.0 | 4.8M | Nov. 2019 |
| v2.0 | 8.8M | Jan. 2020 |
| v3.0 | 21.4M | Dec. 2021 |
| v4.0 | 44.2M | Dec. 2023 |

[JParaCrawl v4.0]
We built a web-based parallel corpus of more than 44M sentence pairs using three different bilingual website lists.

- Crowdsourced websites are more likely to be parallel, although each has fewer parallel sentences
- Improved translation accuracy in a variety of areas, including news, conversation, and scientific papers




## Patent parallel corpus: JaParaPat

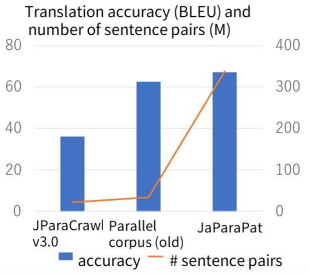



| route | #Docs | #Sents | #Words |
|---|---|---|---|
| Paris | 870K | 180M | 7.4B |
| PCT | 530K | 160M | 6.2B |
| Paris+PCT | 1.40M | 340M | 13.6B |

| 2000-2013Paris | Accu. | #Sents |
|---|---|---|
| Alignment by dictionary | 62.6 | 34M |
| Alignment by translation | 63.4 | 43M |

[JaParaPat]
Based on patent families, we built a patent parallel corpus of more than 300M sentence pairs.

- Covers both the Paris route (priority claims) and the PCT route (international applications)
- Using machine translation for sentence alignment improves the quality and quantity




## References

[1] Makoto Morishita, Katsuki Chousa, Masaaki Nagata, “JParaCrawl v4.0: Building a Large Parallel Corpus with Crowdsourcing,” *proceedings of the 30th a*nnual meeting of Association for Natural Language Processing (NLP-2024), pp. 2330-2335, 2024 (in Japanese).
[2] Masaaki Nagata, Makoto Morishita, Katsuki Chousa, Norihito Yasuda, “JaParaPat: A Large-scale Japanese-English Parallel Patent Application Corpus,” *proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9452-9462, 2024.

## Contact

Masaaki Nagata, Linguistic Intelligence Research Group, Innovative Communication Laboratory