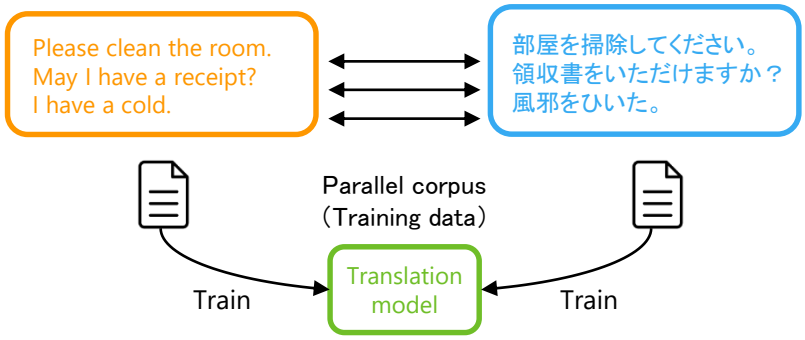
09 Tuning machine translation with small tuning data

Domain adaptation with JParaCrawl, a large parallel corpus

Abstract

Recent machine translation algorithms mainly rely on parallel corpora. However, since the availability of parallel corpora remains limited, only some resource-rich language pairs can benefit from them.
We constructed a parallel corpus for English-Japanese, for which the amount of publicly available parallel corpora is still limited. We constructed the parallel corpus by broadly crawling the web and automatically aligning parallel sentences.
Our collected corpus, called JParaCrawl, amassed over 10 million sentence pairs. JParaCrawl is now freely available online for research purposes.
We show how a neural machine translation model trained with it works as a good pre-trained model for fine-tuning specific domains and achieves good performance even if the target domain data is limited.

Training Machine Translation Model



- Machine Translation (MT) model learns automatically from the parallel sentences (parallel corpus).
  → The amount of parallel sentences is the key to its accuracy.
- We need a large parallel corpus for each domain to make a practical MT system, but such domains are limited.
- Our purpose is to accurately translate low-resource domains, which only have a small number of parallel sentences.

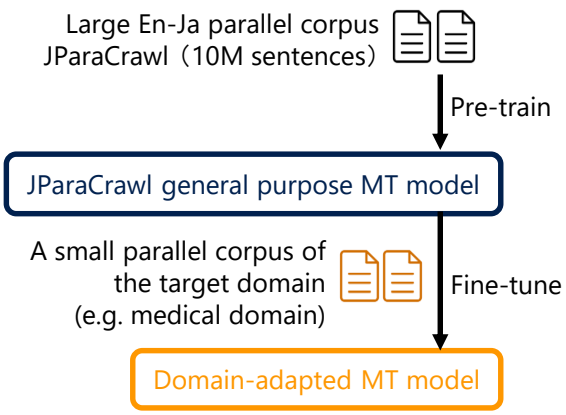
How to create a large En-Ja parallel corpus

Crawl the web (150k websites, 14TB) → Find parallel sentences

- We created a large-scale English-Japanese parallel corpus "JParaCrawl" that contains more than 10M sentences by largely crawling the web and automatically aligning the parallel sentences.
- Until now, freely available parallel corpora are up to 3.0M, so this is more than three times larger than the previous largest one.
- This corpus covers broad domains since it is based on the web.

Our corpus is freely available online for research purposes.
http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

Domain adaptation with small tuning data



- Even if the amount of the target domain parallel corpus is small, we can achieve good performance with the combination of JParaCrawl.
- Fine-tuning only needs small computational cost.

References

[1] M. Morishita, J. Suzuki, M. Nagata, "JParaCrawl: A large scale web-based Japanese-English parallel corpus," *Proc. 12th International Conference on Language Resources and Evaluation (LREC)*, May 2020.

Contact

Makoto Morishita Email: cs-openhouse-ml@hco.ntt.co.jp
Linguistic Intelligence Research Group, Innovative Communication Laboratory

オープンハウス 2020