Domain adaptation with JParaCrawl, a large parallel corpus

09 少量の追加データで作るカスタム機械翻訳

汎用対訳コーパスJParaCrawlを用いた機械翻訳の領域適応

どんな研究

機械翻訳では、対訳コーパスと呼ばれる学習データから自動で翻訳器を学習します。そのため、特定の領域（分野）に特化した翻訳器を作成するためには、その領域の学習データが大量に必要となっていました。この展示では、少量の追加データだけで翻訳器を特定領域に特化させる技術を紹介します。

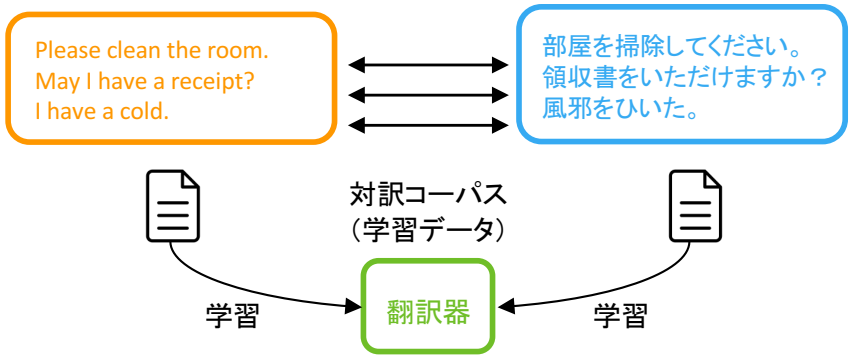
どこが凄い

Webデータを大量に収集し、自動的に対訳になっている文を見つけることで大規模な日本語-英語の学習データを作成しました。この学習データは様々な領域を網羅しているため、これを併用することで少量の学習データだけで特定領域への翻訳器を特化させることが可能になりました。

めざす未来

本技術を用いることで、これまで学習データが乏しかったため翻訳精度が低かった領域に対しても、少量の追加学習データで翻訳精度を飛躍的に向上させることが可能になります。将来的には、どの領域に対しても高精度な機械翻訳の実現をめざします。

機械翻訳器の学習



- 機械翻訳器は、同じ意味の2言語文対（対訳コーパス）を大量に用意することで自動的に学習
  → 対訳コーパスの量が精度に大きく影響
- 実用化に十分な量の対訳コーパスがある領域は限定的（旅行・特許等）
- 新たにデータを大量に集めるのは大変
  → 何らかの工夫が必要
- 少量しか対訳コーパスが存在しない領域も正しく翻訳できるようにしたい

大規模日英対訳コーパスの作成

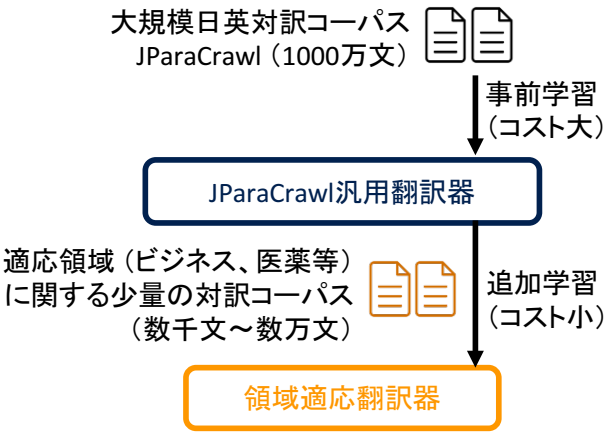
Webデータ収集
（15万ウェブサイト, 14TB）

日英対訳文抽出

- Webデータを大規模に収集し、日英対訳文を自動的に抽出することにより、1000万文を超える大規模日英対訳コーパス“JParaCrawl”を作成
- これまで無償で公開されている日英対訳コーパスは高々300万文だったので、従来の3倍以上の大きさ
- 本対訳コーパスはWebをもとにしているので、様々な領域を網羅しているという特長あり

本対訳コーパスは研究目的利用に限り以下のURLから無償公開しています。
http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

少量追加データでの領域適応



- 適応領域に関する対訳コーパスが少量しか用意できなくても、様々な分野を含んでいる大規模日英対訳コーパスJParaCrawl を組み合わせることで、高い翻訳精度を達成
- 追加学習は短時間ですみ、学習にかかるコストも低い

関連文献

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

連 絡 先

森下 睦（Makoto Morishita） 協創情報研究部 言語知能研究グループ
Email: cs-openhouse-ml@hco.ntt.co.jp

オープンハウス 2020