Large-scale parallel corpus construction technology

07 機械翻訳モデル訓練のための大規模対訳コーパス構築技術

世界中から対訳を探します

どんな研究	機械翻訳システムは、訓練データとして対訳文対（お互いに翻訳である文の対）を大量に必要とします。われわれはインターネット（Web）や特許出願アーカイブに点在する対訳データを収集して対訳コーパス（対訳データベース）を構築する技術の研究を行っています。
どこが凄い	Web対訳コーパスJParaCrawlはクラウドソーシングを活用することで多くの対訳データをWebから効率的に収集し構築しました。特許対訳コーパスJaParaPatは、データ収集とモデル訓練を交互に繰り返すことで対訳データの品質を高めました。どちらも公開されている中では世界最大の日英対訳コーパスです。
めざす未来	専門用語が豊富な医療や金融など特定の分野、中国語と日本語など特定の言語対について高品質な対訳データを自動的に収集する技術をさらに高め、お客様のニーズに応じて独自にカスタマイズできる機械翻訳技術の実現をめざします。

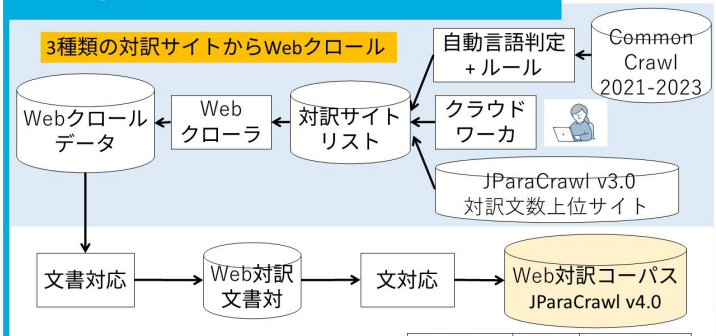
機械翻訳と対訳データ

機械翻訳の精度は対訳データの量と質で決まる

言語はコミュニケーションの手段である

インターネット(Web)や特許を対象に大量かつ高品質な対訳コーパスを構築

Web対訳コーパス: JParaCrawl v4.0

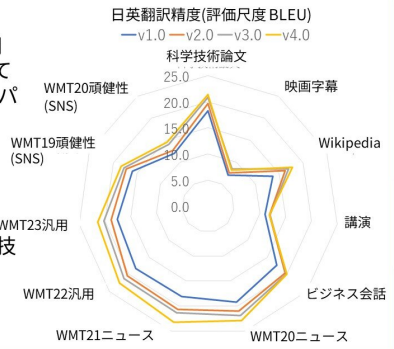


対訳サイト探索法	件数	対訳存在	文数
CommonCrawl	5.0万	35%	22.8M
クラウドワーカ	2.1万	85%	16.9M
v3.0上位サイト	2.4万	90%	6.9M

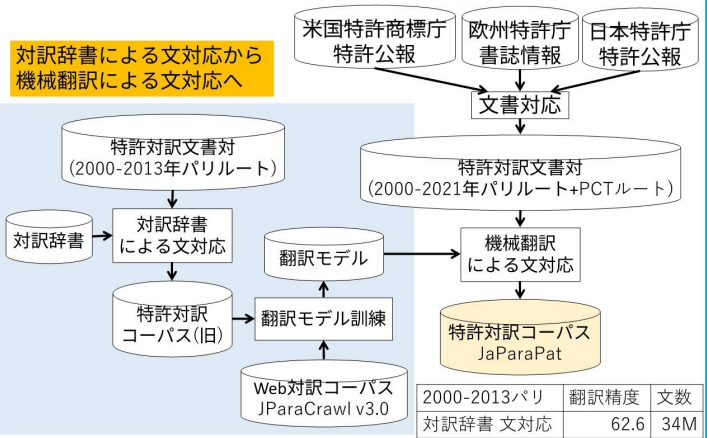
バージョン	文数	作成時期
v1.0	4.8M	2019年11月
v2.0	8.8M	2020年 1月
v3.0	21.4M	2021年12月
v4.0	44.2M	2023年12月

[JParaCrawl v4.0]
3種類の対訳サイトリストを使用することで、これまでとあわせて4400万文対以上のWeb対訳コーパスを構築

- クラウドソーシングによる対訳Webサイトは個々の対訳抽出数が少ないが、対訳抽出の成功率が高い
- ニュース、SNS、会話、科学技術論文など様々な領域で翻訳精度が向上



特許対訳コーパス: JaParaPat

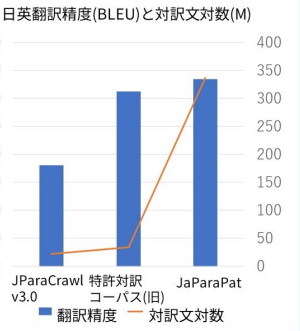


2000-2013パリ	翻訳精度	文数
対訳辞書 文対応	62.6	34M
機械翻訳 文対応	63.4	43M

ルート	文書数	文数	単語数
パリ	87万	1.8億	74億
PCT	53万	1.6億	62億
パリ+PCT	140万	3.4億	136億

[JaParaPat]
書誌情報(パテントファミリー)に基づいて2000-2021年の特許出願から3億文対以上の特許対訳コーパスを構築

- パリルート(優先権主張)とPCTルート(国際出願)を網羅的に探索
- 機械翻訳による文対応を用いることで対訳辞書による文対応に比べて対訳コーパスの質・量ともに向上



関連文献

[1] 森下睦, 帖佐克己, 永田昌明, “JParaCrawl v4.0: クラウドソーシングを併用した大規模対訳コーパスの構築,” *言語処理学会第30回年次大会*, pp. 2330-2335, 2024.

[2] 永田昌明, 森下睦, 帖佐克己, 安田宜仁, “JaParaPat: 大規模日英特許対訳コーパス,” *言語処理学会第30回年次大会*, pp. 2367-2372, 2024.

連絡先

永田 昌明（Masaaki Nagata）協創情報研究部 言語知能研究グループ