JParaCrawl: 大規模Webベース日英対訳コーパス

NTT コミュニケーション科学基礎研究所
森下 睦, 鈴木 潤, 永田 昌明

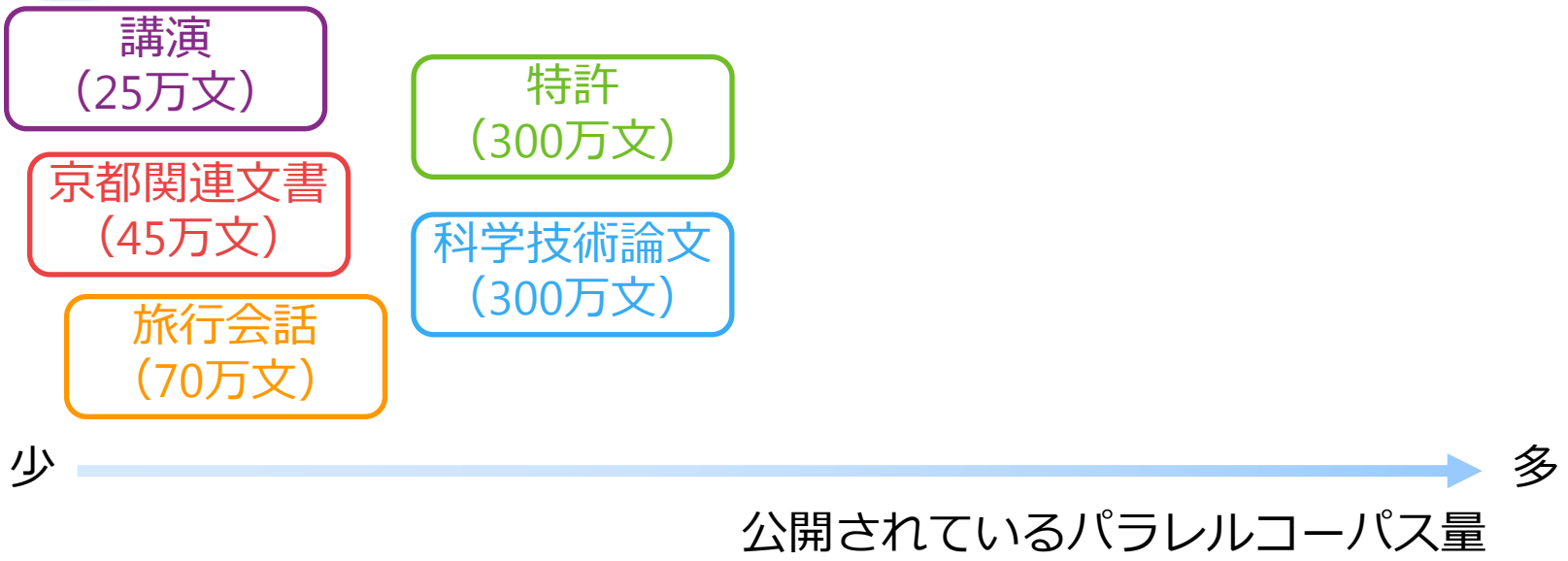
JParaCrawl

- 約1000万文対の大規模日英対訳コーパス
  - Webを大規模にクロールすることで作成
  - Webをもとにしているので，様々な分野を含んでいる
- 研究目的に限り無償で公開
  - 無償公開されている中では，最大の日英対訳コーパス
- JParaCrawlで学習したNMTモデルも公開
  - 適応先ドメインでfine-tuningを行うことで，
    短時間でドメイン特化NMTモデルを作成可能

大規模日英対訳コーパス
JParaCrawl

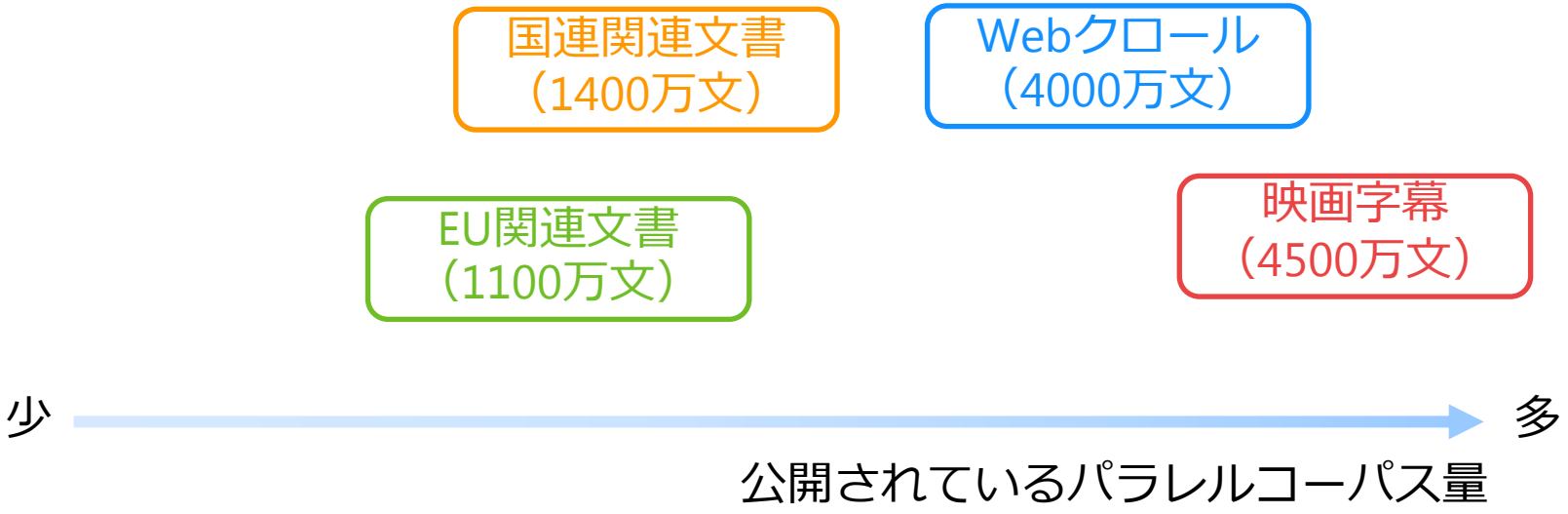
日英対訳コーパスの現状

- 無償で公開されている対訳コーパスは
  少量かつ分野も限定的

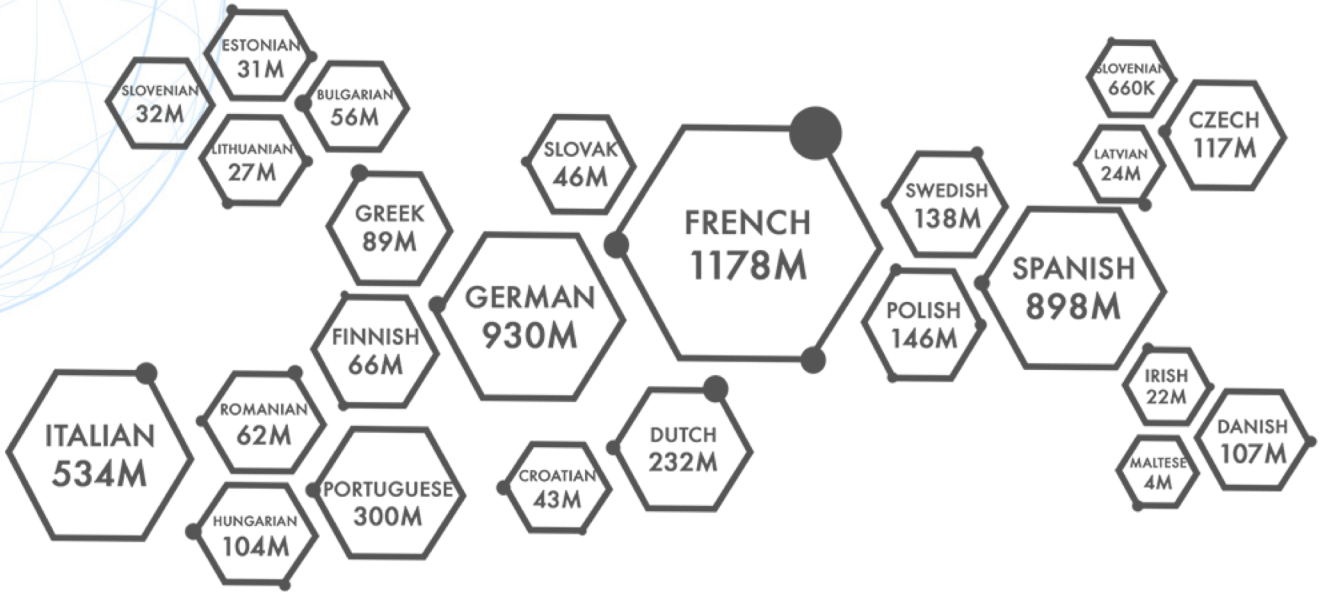


仏英対訳コーパスの現状

- 日英対訳コーパスと比べて桁が違う
  → 数千万文対訳コーパスがあると，
  特定分野については人手と同等の翻訳精度が
  達成できると言われている



ParaCrawl



- ヨーロッパ言語-英語間の
  大規模な対訳コーパスを作成するプロジェクト
- EUからの助成を受けている
  - このまま待っていても（おそらく）日英は出てこない

本プロジェクトの目的

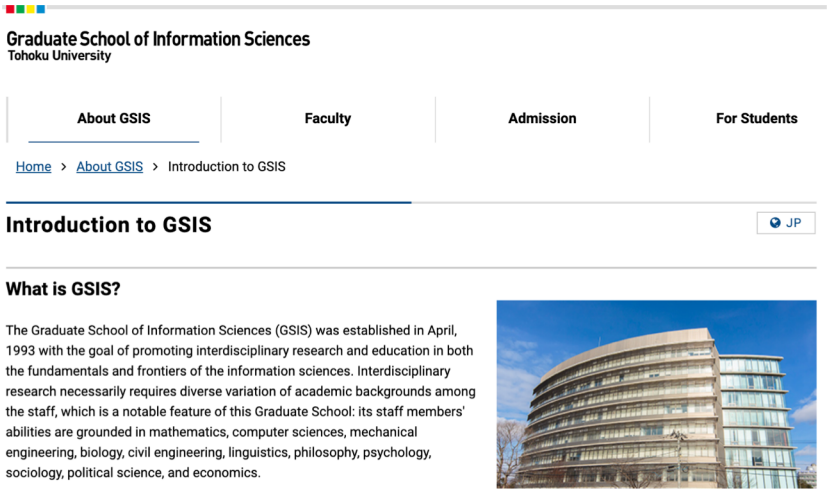
- このままだとヨーロッパにおいていかれる!?
  - 機械翻訳業界で日英をメジャーな言語対にしたい
  - 日英で論文を通しやすくしたい

- 日英大規模対訳コーパス**JParaCrawl**を作成

Webからの対訳データ収集

- 同一ドメイン上に対訳が存在するWebサイトから対訳文を抽出
  - 手法はほぼ本家ParaCrawlを踏襲

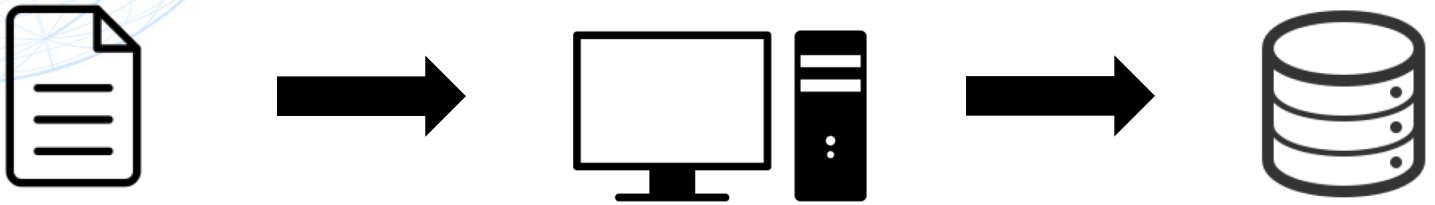
https://www.is.tohoku.ac.jp/jp/introduction/outline.html



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

収集方針

1. CommonCrawl上に存在する全てのテキストを解析
   - 各ドメインについて，言語別のデータ量を得る



CommonCrawl　　CLD2を用いた言語検出　　ドメイン別言語データ量

ドメイン名	英語 [KB]	日本語 [KB]
xxx.jp	1000	2000
yyy.com	45000	100
zzz.co.jp	3000	2500
…	…	…

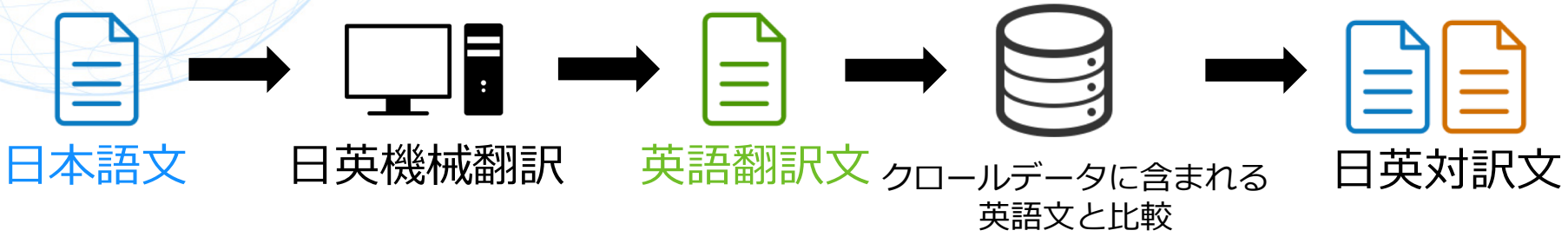
収集方針

2. クロール候補ドメインの抽出
   - 英日テキストデータの比に着目
   - 同等の比をもつドメイン15万件を抽出

3. 候補Webサイトをクロール
   - AWS EC2インスタンス50台を使って約2ヶ月クロール
   - 圧縮した状態で14.4TBのデータを得られた

収集方針

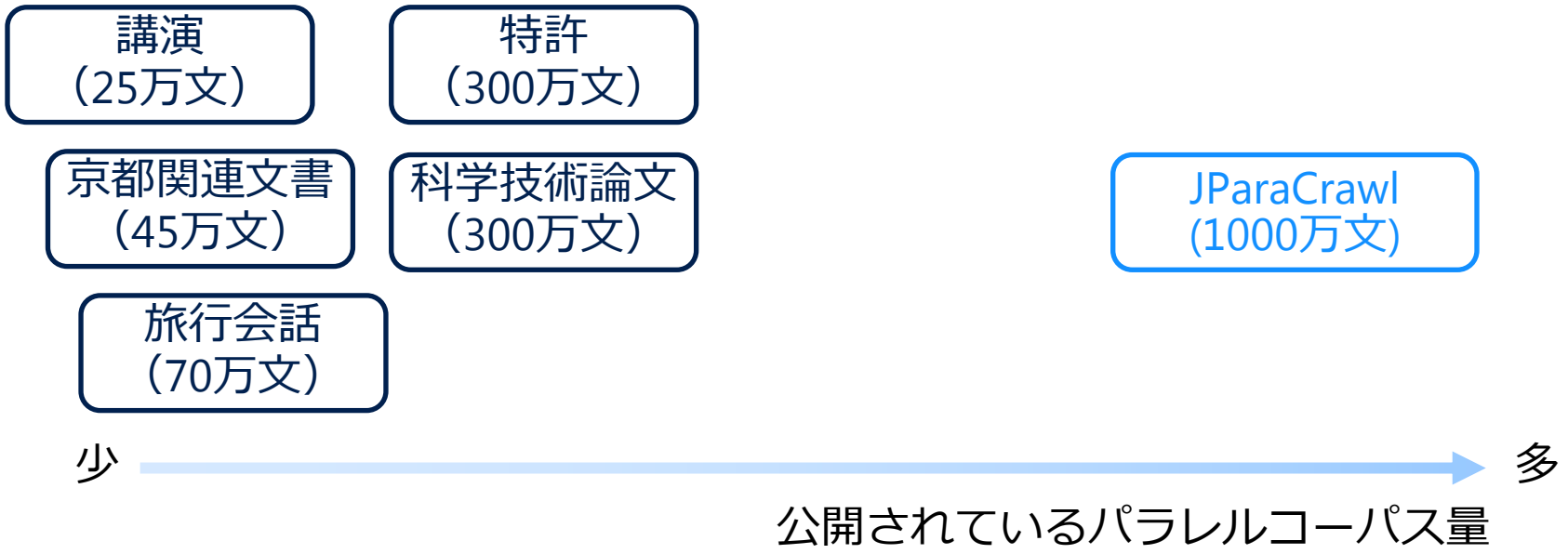
4. 収集したデータから対訳文を抽出
   – 日英翻訳した英文と類似している英文をデータから探索



5. ノイジーな対訳文をフィルタリング
   – ルール，辞書，言語モデル等に基づいてノイズを除去

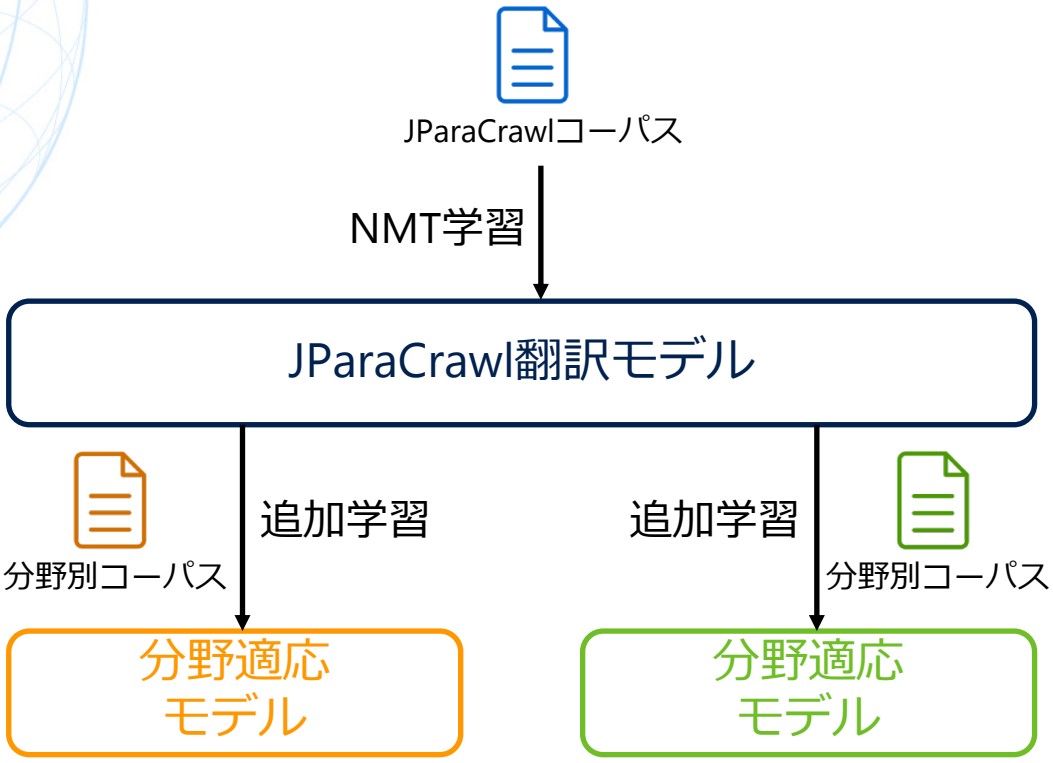
JParaCrawl

- 1000万文以上のクリーンな対訳文を抽出
  - 現在無償で公開されているものと比較すると最も大きい日英対訳コーパス
  - Webをもとにしているので幅広い分野をカバー



実験

JParaCrawlからの分野適応



- JParaCrawlから特定分野へ短時間で適応可能か確認
  → Pre-trained モデルとして使用可能？

実験設定

- **比較対象:**
  - 分野別データのみで最初から学習したモデル

- **翻訳モデル:** Transformer

- **分野別学習・テストセット / ハイパーパラメータ:**

- 科学技術論文 (ASPEC)	約300万文 / big
- 映画字幕 (JESC)	約280万文 / big
- 京都関連Wikipedia記事 (KFTT)	約44万文 / base
- TEDトーク (IWSLT)	約22万文 / small

実験結果

	英日翻訳			日英翻訳		
評価データ	分野別モデル	JParaCrawl	＋ Fine-tuning	分野別モデル	JParaCrawl	＋ Fine-tuning
ASPEC	44.3	26.5 (-17.8)	43.9 (-0.5)	28.7	19.7 (-9.0)	29.2 (+0.5)
JESC	14.5	6.5 (-8.0)	13.9 (-0.6)	17.8	7.5 (-10.3)	17.5 (-0.3)
KFTT	31.8	18.9 (-12.9)	33.2 (+1.4)	23.4	16.2 (-7.2)	25.9 (+2.5)
IWSLT	11.1	12.6 (+1.5)	14.5 (+3.4)	13.7	11.9 (-1.8)	17.2 (+3.5)

- Fine-tuningを行うことで，分野別モデルと
  ほぼ同等もしくはそれ以上の精度を達成
- 特に分野別学習データが小さい分野では
  大幅な精度向上を達成

JParaCrawlの公開と著作権法

著作権法の問題

- 日本の著作権法が2019年1月に改正
  - 改正前
    - 第三者の著作物が含まれているWebデータは配布できない
    - フェアユースの規定は無い
  - 今回の法改正は，近年の機械学習などの流れを強く意識
    - コンピュータ上での著作物の処理に関する規定がかなり緩くなる

改正著作権法

改正著作権法 新30条の4

著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情の享受を目的としない場合には、その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。

- 第三者の著作物を集めた対訳コーパスは配布可能？
  - 対訳コーパス = 思想又は感情の享受を目的としない
  - 「利用することができる」
    - 翻訳モデルの学習等
    - 第三者への学習データの配布も含む
  - （私達の解釈では）配布可能

コーパスと事前学習モデルの公開

- コーパスとJParaCrawlの事前学習モデルを公開中!
  http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

- Fine-tuningの使用例スクリプトも公開中
  - 機械翻訳研究への参入障壁が下がってほしいという思い
  - ドキュメントをちゃんと書いた（つもりです）
  - https://github.com/MorinoseiMorizo/jparacrawl-finetune

宣伝

- WMT20 ニュース翻訳タスク に日英タスクが追加！
  - 学習データ，Devセットも公開済み
  - http://www.statmt.org/wmt20/

- 皆さんのご参加をお待ちしております!!!

EMNLP 2020
FIFTH CONFERENCE ON
MACHINE TRANSLATION (WMT20)

November 11-12, 2020
Punta Cana, Dominican Republic

Shared Task: Machine Translation of News

[HOME]
TRANSLATION TASKS: [NEWS] [LIFELONG LEARNING] [ROBUSTNESS] [QUALITY ESTIMATION] [UNSUP AND VERY LOW RES]

The recurring translation task of the WMT workshops focuses on news text. For this year the language pairs are:

- Chinese-English
- Czech-English (this year **both directions again**)
- French-German
- German-English
- Inuktitut-English
- Japanese-English
- Polish-English
- Russian-English
- Tamil-English

We provide parallel corpora for all languages as training data, and additional resources for download.

次のWMT開催地 Punta Cana はビーチで有名らしい

お伝えしたいメッセージ

- 1000万文を超える
  大規模日英対訳コーパスJParaCrawlを作成
  → 研究目的に限り無償公開中!
  → http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

- より多くの方に機械翻訳研究に興味をもってほしい
  → 今回提供したスクリプトなどがきっかけになる？

- 日英翻訳がより研究業界でメジャーになってほしい
  → WMT20に日英タスクを追加できたので,
  皆さんのご参加をお待ちしております！

END