Expanding the Applicability of Machine Translation

機械翻訳の適応先拡張

森下 睦

NTT コミュニケーション科学基礎研究所

2022/09/28 AAMTセミナー

自己紹介

森下 睦 (もりした まこと)

2015-2017: 奈良先端科学技術大学院大学, 修士 (工学)

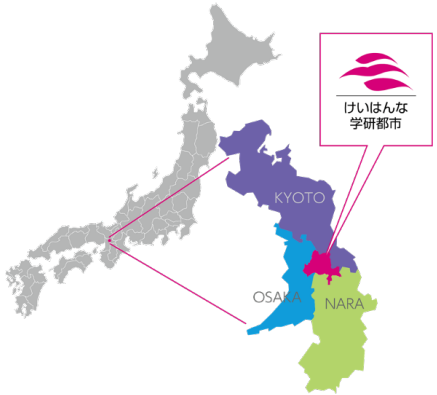
2017-: NTT コミュニケーション科学基礎研究所

2020-2022: 東北大学, 博士 (情報科学)

→ 博士論文: “Expanding the Applicability of Machine Translation”

研究領域: 機械翻訳

京阪奈生まれ、京阪奈育ち、京阪奈勤務



AAMT長尾賞 学生奨励賞受賞

- 東北大学での博士論文に対して AAMT長尾賞学生奨励賞をいただきました
- 選考委員の方々をはじめ関係者の皆様に、この場をお借りして御礼申し上げます

博士論文の構成

1. 日英大規模対訳コーパスの構築
2. 低資源ドメインに対するドメイン適応
3. ドキュメント全体を考慮した機械翻訳モデル

博士論文の構成

1. 日英大規模対訳コーパスの構築
2. 低資源ドメインに対するドメイン適応
3. ドキュメント全体を考慮した機械翻訳モデル

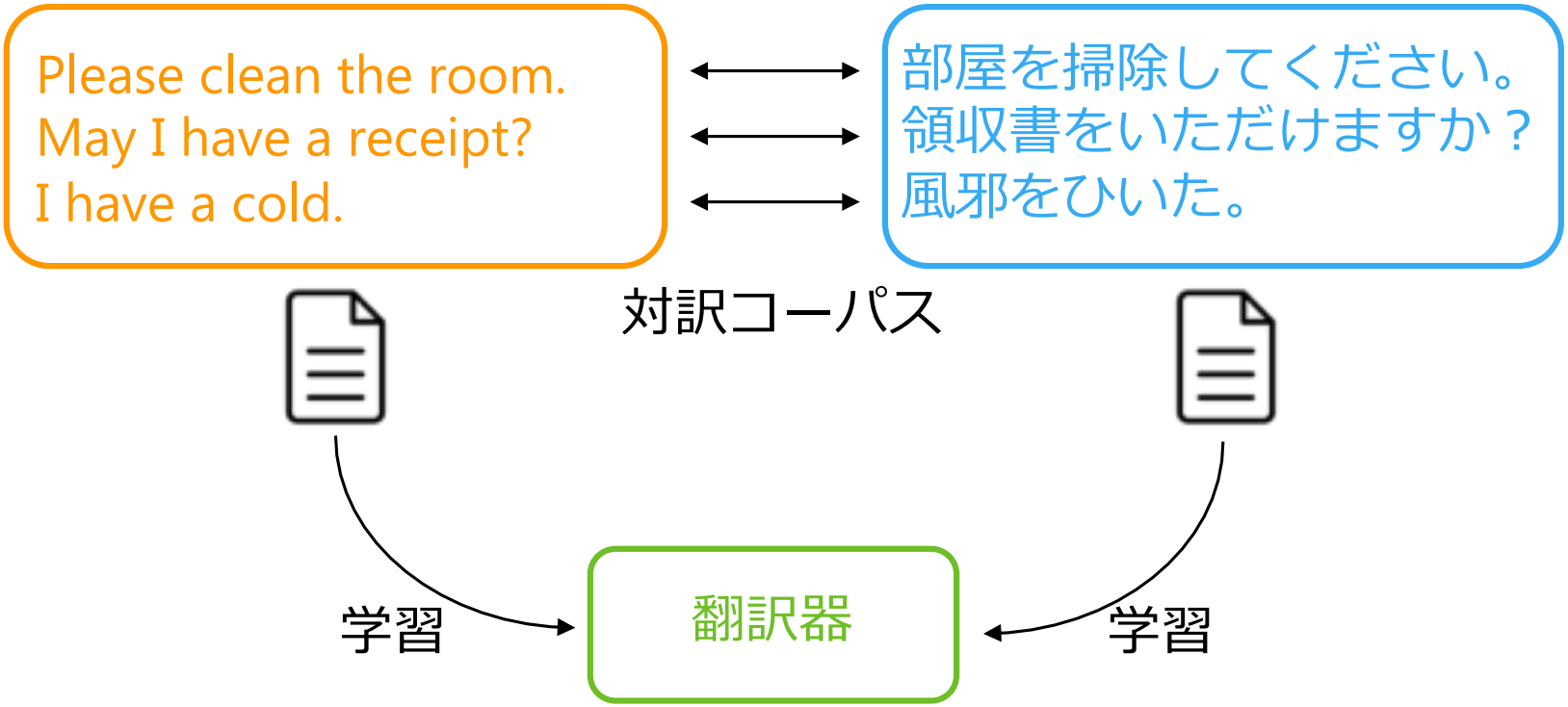
日英大規模対訳コーパスの構築

概要

- **課題**
  - 研究用に一般公開されている日英データの不足
  - 日英学習データが小さいことによる全体的な翻訳精度の問題
- **解決方法**
  - Webをクロールし大規模日英対訳データを構築
- **本研究のインパクト**
  - 世界最大級の日英対訳データを構築し公開
  - 様々な分野で日英翻訳精度が大幅に向上
  - 世界の日英翻訳研究のデファクトスタンダード

    学習データとなった

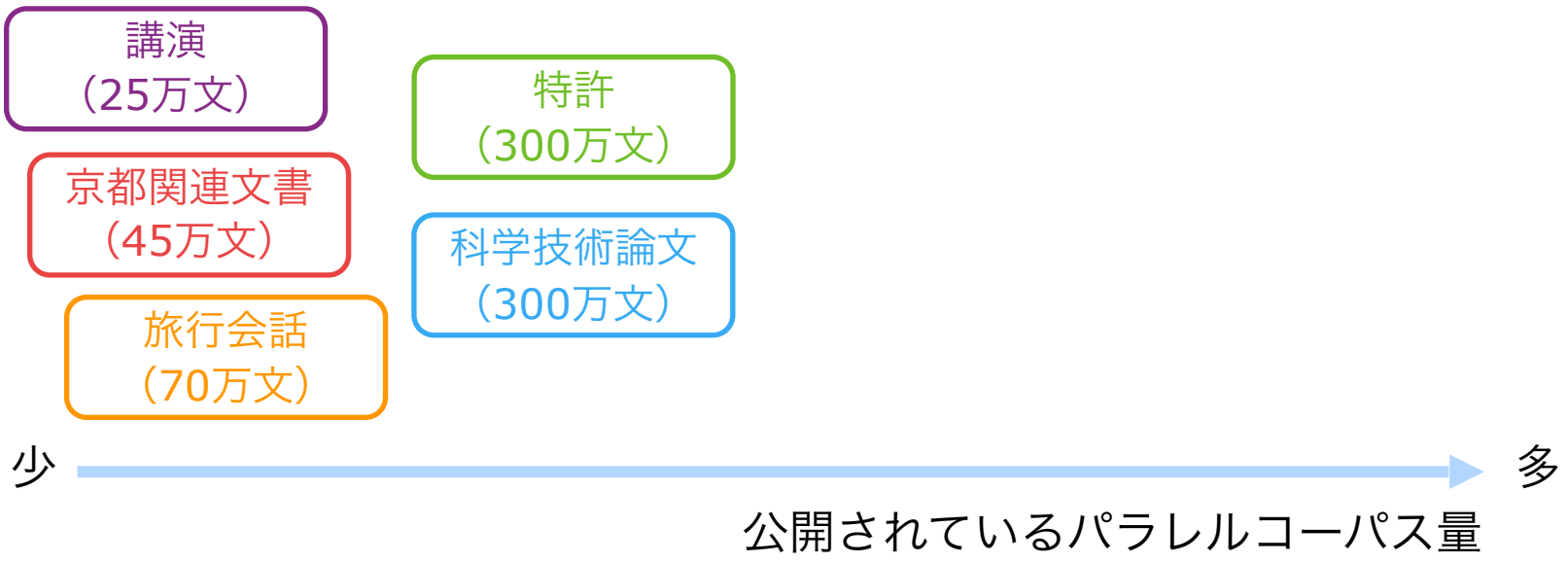
機械翻訳の学習



- 対訳データから翻訳モデルを自動的に学習

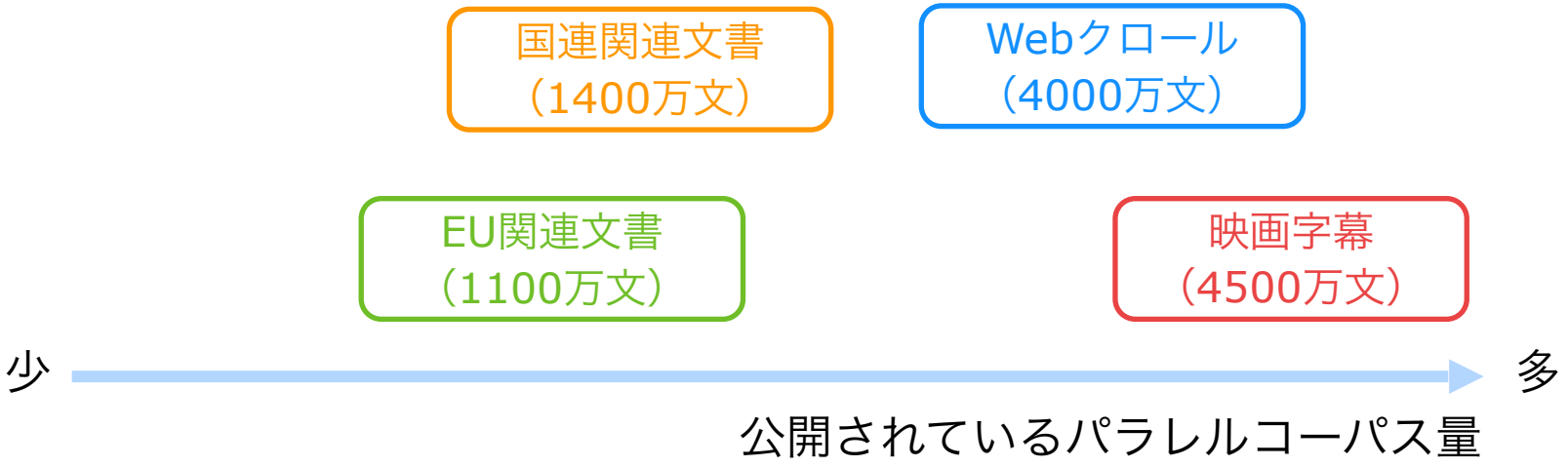
日英対訳コーパスのこれまで

- 無償公開されている対訳コーパスは少量かつ分野も限定的

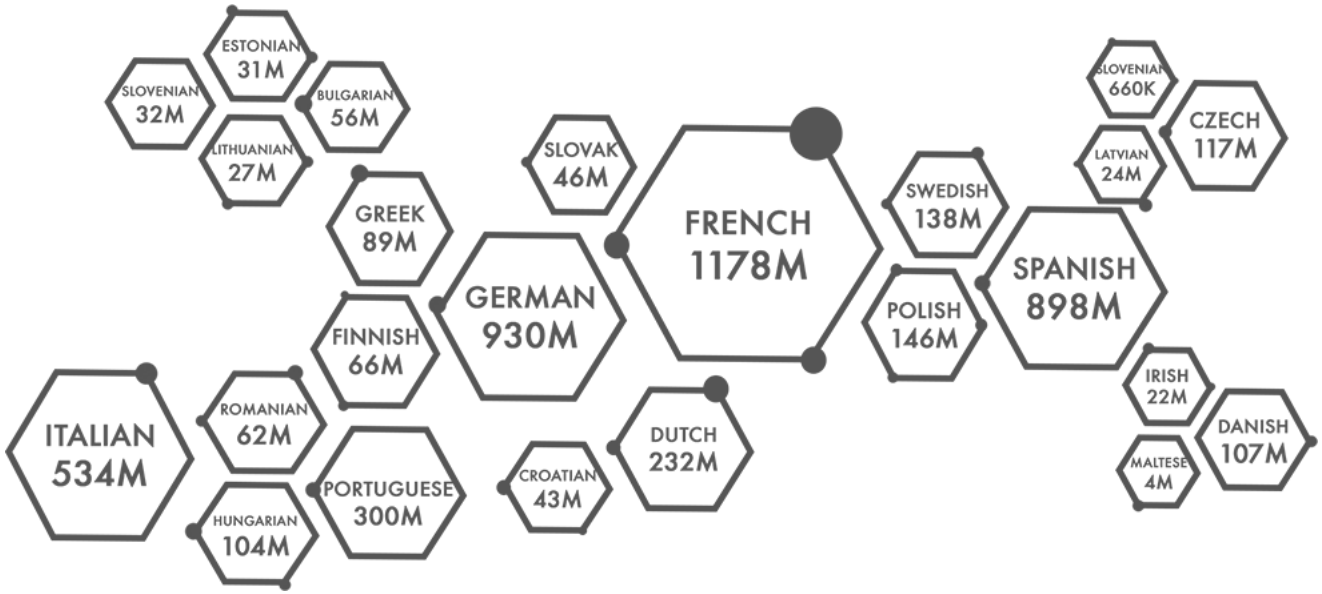


仏英対訳コーパスの現状

- 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

機械翻訳研究における「メジャー言語対」

- 機械翻訳研究には、資源が豊富で、頻繁に実験に使われる「メジャー言語対」が存在する（と思う）
  - 例: フランス語-英語、ドイツ語-英語、中国語-英語
  - こういった言語対で実験していると、ケチを付けられることは少ない（ように思う）
  - ゆえに、以前は私も仏英、独英で実験していた
  - ただ、読めない言語で実験するのは不利だし、日本語話者としては本当は日本語に取り組みたい・・・

本プロジェクトの目的

- 日英翻訳の精度を底上げしたい
- 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

なぜこれまで大規模日英データがなかった？

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。

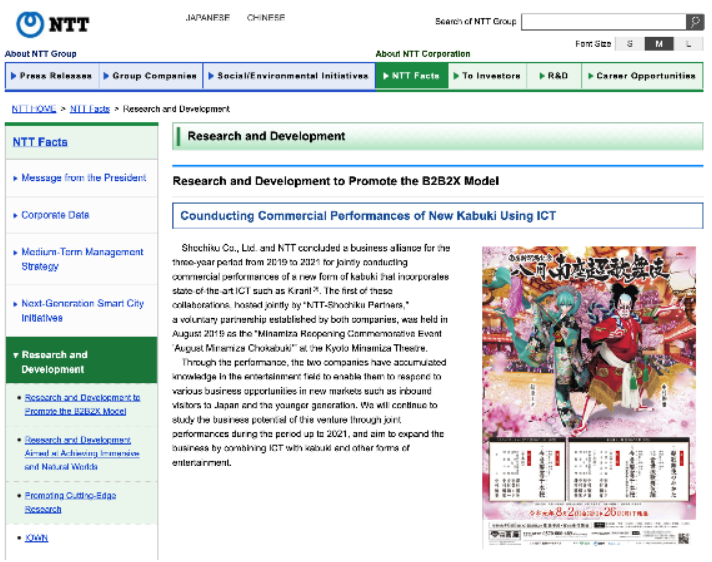
第三者の著作物を集めた対訳コーパスは配布可能？

- 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

対訳コーパスの構築

Webからの対訳文収集

www.ntt.co.jp/about/r_d.html



www.ntt.co.jp/about_e/r_d.html

- Webは対訳文が眠っている鉱山 
- 対訳文が含まれているWebサイト (ドメイン) を見つけて、自動で対訳文を抽出

どのように対訳文が含まれているドメインを見つけるか

大量のテキストデータ

✓ aaa.co.jp En: 5MB, Ja: 10MB

✖ bbb.com En: 10MB, Ja: 0.1MB, Fr: 0.3MB

✓ ccc.ac.jp En: 1MB, Ja: 2MB

⋮

- CommonCrawl を分析して、

  各ドメインにおける各言語のデータサイズを得る

  → 言語検出ツール (CLD2) を用いる

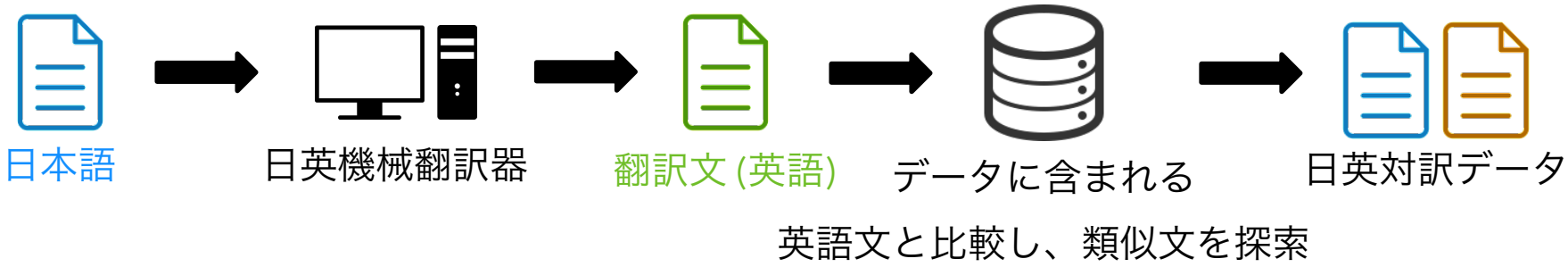
- 日英のデータ量が比較的同等のドメインを列挙

  → 今回は10万ドメインを列挙し、ドメイン全体をクロール

文対応付け

クロールされたデータから対訳文を抽出

- 機械翻訳器を用いて英文に翻訳し類似している英文を探す



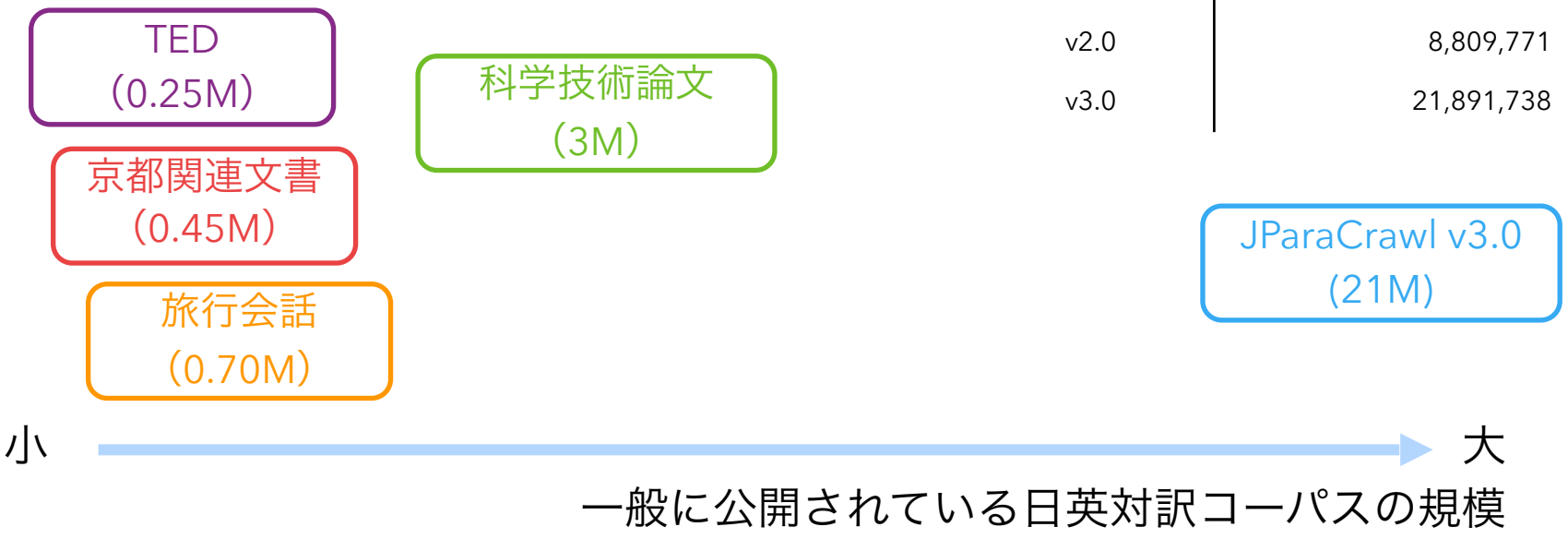
ノイジーな対訳文を取り除く

- 人手で書かれたルール、辞書、言語モデル等を用いてノイズを除去

JParaCrawl

- 最新版の v3.0 では2100万文を超えるコーパスに
  - 一般に無償公開されている日英対訳コーパスの中では最大規模

バージョン	ユニークな対訳文数
v1.0	4,817,172
v2.0	8,809,771
v3.0	21,891,738



実験

翻訳実験

- 様々な分野の評価データを用いて実験

  → JParaCrawlが各分野でどの程度の精度が出せるのか確認

評価データ	領域
ASPEC	科学技術論文
JESC	映画字幕
KFTT	Wikipedia記事
TED	TEDトーク
ビジネスシーン対訳コーパス	会話文
WMT20 ニュースタスク En-Ja	ニュース
WMT20 ニュースタスク Ja-En	ニュース
WMT21 ニュースタスク En-Ja	ニュース
WMT21 ニュースタスク Ja-En	ニュース
WMT19 頑健性タスク En-Ja	SNS
WMT19 頑健性タスク Ja-En	SNS
WMT20 頑健性タスク Set1 En-Ja	Wikipediaコメント
WMT20 頑健性タスク Set2 En-Ja	SNS
WMT20 頑健性タスク Set2 Ja-En	SNS
IWSLT21 同時通訳タスク En-Ja	TEDトーク

実験結果 (英日翻訳)

評価データ	領域	v1.0	v2.0	v3.0	v3.0-v2.0
ASPEC	科学技術論文	24.7	26.5	**26.8**	+0.3
JESC	映画字幕	**6.6**	6.5	6.5	0.0
KFTT	Wikipedia記事	17.1	**18.9**	18.1	-0.8
TED	TEDトーク	11.5	12.6	**13.1**	+0.5
ビジネスシーン対訳コーパス	会話文	12.4	13.5	**13.9**	+0.4
WMT20 ニュースタスク En-Ja	ニュース	20.7	21.9	**23.5**	+1.6
WMT20 ニュースタスク Ja-En	ニュース	20.1	22.8	**23.5**	+0.7
WMT21 ニュースタスク En-Ja	ニュース	21.1	21.8	**25.0**	+3.2
WMT21 ニュースタスク Ja-En	ニュース	19.6	21.5	**22.4**	+0.9
WMT19 頑健性タスク En-Ja	SNS	12.4	12.5	**14.4**	+1.9
WMT19 頑健性タスク Ja-En	SNS	11.5	12.3	**12.8**	+0.5
WMT20 頑健性タスク Set1 En-Ja	Wikipediaコメント	15.2	15.8	**18.7**	+2.9
WMT20 頑健性タスク Set2 En-Ja	SNS	12.7	13.0	**14.8**	+1.8
WMT20 頑健性タスク Set2 Ja-En	SNS	7.9	8.2	**8.6**	+0.4
IWSLT21 同時通訳タスク En-Ja	TEDトーク	12.5	13.3	**14.5**	+1.2

多くの評価データで高い翻訳精度を達成

実験結果 (英日翻訳)

評価データ	領域	v1.0	v2.0	v3.0	v3.0-v2.0
ASPEC	科学技術論文	24.7	26.5	**26.8**	+0.3
JESC	映画字幕	**6.6**	6.5	6.5	0.0
KFTT	Wikipedia記事	17.1	**18.9**	18.1	-0.8
TED	TEDトーク	11.5	12.6	**13.1**	+0.5
ビジネスシーン対訳コーパス	会話文	12.4	13.5	**13.9**	+0.4
WMT20 ニュースタスク En-Ja	ニュース	20.7	21.9	**23.5**	+1.6
WMT20 ニュースタスク Ja-En	ニュース	20.1	22.8	**23.5**	+0.7
WMT21 ニュースタスク En-Ja	ニュース	21.1	21.8	**25.0**	+3.2
WMT21 ニュースタスク Ja-En	ニュース	19.6	21.5	**22.4**	+0.9
WMT19 頑健性タスク En-Ja	SNS	12.4	12.5	**14.4**	+1.9
WMT19 頑健性タスク Ja-En	SNS	11.5	12.3	**12.8**	+0.5
WMT20 頑健性タスク Set1 En-Ja	Wikipediaコメント	15.2	15.8	**18.7**	+2.9
WMT20 頑健性タスク Set2 En-Ja	SNS	12.7	13.0	**14.8**	+1.8
WMT20 頑健性タスク Set2 Ja-En	SNS	7.9	8.2	**8.6**	+0.4
IWSLT21 同時通訳タスク En-Ja	TEDトーク	12.5	13.3	**14.5**	+1.2

多くの評価データで高い翻訳精度を達成

→ 特にニュース分野の伸びが大きい

→ データに最新の語彙を多数含んでいることによる伸びか

翻訳例

原文	院内に「濃厚接触者」はいませんが、接触者全員に PCR 検査を実施し、 女性が関係した病棟などを閉鎖して徹底的に消毒するということです。
参照訳	There are no known “close contacts” in the hospital, but all contacts will be subjected to PCR tests, and the wards and other areas where the women had been will be closed and thoroughly disinfected.
JParaCrawl v1.0	There is no “strong contact person” in the hospital, but a PCR test will be conducted for all the contacts, and women will close the wards and thoroughly disinfect them.
JParaCrawl v2.0	Although there is no “strong contact person” in the hospital, PCR tests will be performed on all contact persons, and the wards related to women will be closed and thoroughly disinfected.
JParaCrawl v3.0	There are no “close contacts” in the hospital, but PCR tests will be conducted for all contacts, and the wards related to women will be closed and thoroughly disinfected.

- JParaCrawl v3.0は2021年のWebデータをもとに構築されている
- ゆえに、v3.0をもとにしたモデルは

  近年頻出する「濃厚接触者」を比較的正しく翻訳できている

日英・英日機械翻訳タスクの主催

機械翻訳シェアードタスクWMT

- 機械翻訳の研究分野において

  最も有名かつ競争の激しいコンペティション

  - 2006年より毎年開催
  - 機械翻訳の学習用データが与えられ、

    参加者はそれをもとに高い精度が出る翻訳器を学習
  - 毎年世界中の企業・大学等が参加
    - 多数の機械翻訳研究者が毎年結果に注目
  - これまで日英翻訳は対象外
    - 主に独英・中英などが主戦場

日本語タスクの主催

- 世界的な機械翻訳シェアードタスクWMTで使用された言語対、
  データはメジャーな研究対象として認識され、
  その後の研究でも使用される傾向にある
  → 日英翻訳をよりメジャーな言語対にしたい
  → 多くの人が日英を対象に実験を行ってほしい
  → 世界の協力を得て日英翻訳の技術改良を進めたい
- 主催者と交渉し2020年から日英タスクを開催

開催結果

- これまで日本人しか取り組んでいなかった日英翻訳に
  世界中から多数のチームが取り組む
  →日英言語対のメジャー言語化に貢献

人手による翻訳評価結果

2020年

English→Japanese

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z
53.9	−0.132	SJTU-NICT
52.8	−0.164	Online-G

2021年

English→Japanese

Rank	Ave.	Ave. z	System
1–2	86.4	0.430	Facebook-AI
1–2	85.3	0.314	HUMAN-A
3–5	84.2	0.266	Online-W
3–5	81.3	0.168	WeChat-AI
3–5	82.6	0.148	NiuTrans
6–8	77.8	0.017	HW-TSC
6–8	71.8	-0.042	MiSS
8–13	78.5	-0.051	Online-Y
6–10	77.8	-0.067	BUPT_rush
8–13	70.9	-0.129	Online-A
9–13	67.4	-0.184	Online-B
9–14	74.2	-0.284	ephemeraler
9–14	72.5	-0.339	capitalmarvel
12–14	70.1	-0.373	movelikeajaguar
15–16	63.5	-0.440	Illini
15–16	65.7	-0.541	Online-G

まとめ

- Webを大規模にクロールすることで、

  2000万文対を超える大規模な日英対訳コーパスを作成

- 研究目的に限り無償公開
  - 一般に無償公開されている中では、最大級の日英対訳コーパス
- **研究のインパクト**
  - 日英機械翻訳の精度底上げに貢献
  - 日英言語対をよりメジャーな言語対に
    - 日英機械翻訳に世界中の多くの人が取り組むように

END