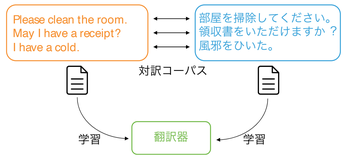
PT4-11

JParaCrawl v3.0: 大規模日英対訳コーパス

森下 睦[1, 2], 帖佐 克己[1], 鈴木 潤[2], 永田 昌明[1]
[1]NTT コミュニケーション科学基礎研究所 , [2] 東北大学

概要

現在の機械翻訳器は主に対訳コーパスから翻訳モデルを学習
→ 対訳コーパスの量・品質が翻訳精度に大きく影響
→ しかし、公開されている日英の対訳コーパスは限定的で、
日英機械翻訳研究の大きな課題となっていた



貢献

Web をもとに大規模な日英対訳コーパスを構築
→ 2100 万文を超え、JParaCrawl v2.0 の倍以上の規模
→ 本コーパスは研究目的利用に限り無償公開
→ WMT2022 機械翻訳タスクにも学習データとして提供予定

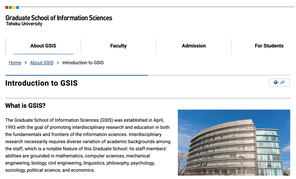
http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

Web からの対訳文収集

収集手順

1. 同一ドメイン上に対訳が存在する Web サイトを CommonCrawl をもとに探索
   → 日英のデータ量が近い 10 万ドメインを列挙
2. 収集対象ドメイン全体をクロール
3. 得られたデータから対訳文を抽出
   → 機械翻訳ベースの対訳抽出法 bleualign を使用
4. ルール、辞書、言語モデル等を用いて
   ノイジーな文をクリーニング

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



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

収集結果

これまでのデータと合わせて
2100 万文以上の対訳コーパスを構築
→ JParaCrawl v3.0 として公開
→ v2.0 と比較して 2 倍以上の規模

バージョン	対訳文数	作成時期
v1.0	4,817,172	2019年11月
v2.0	8,809,771	2020年1月
v3.0	21,481,513	2021年12月

収集対象ドメインのうち 7 割は v2.0 では収集対象外
→ 前回の収集から約 2 年が経過しており、
その間に対訳が存在する Web サイトが変化した

実験

翻訳精度

		英日翻訳				日英翻訳			
評価データ	領域	v1.0	v2.0	v3.0	v3.0-v2.0	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	18.3	19.7	**20.8**	+1.1
JESC	映画字幕	**6.6**	6.5	6.5	0.0	7.0	7.5	**8.4**	+0.9
KFTT	Wikipedia記事	17.1	**18.9**	18.1	-0.8	13.7	16.2	**17.0**	+0.8
TED	TEDトーク	11.5	12.6	**13.1**	+0.5	11.0	11.9	**12.0**	+0.1
ビジネスシーン対訳コーパス	会話文	12.4	13.5	**13.9**	+0.4	17.4	19.6	**19.9**	+0.3
WMT20 ニュースタスク En-Ja	ニュース	20.7	21.9	**23.5**	+1.6	21.3	23.3	**23.9**	+0.6
WMT20 ニュースタスク Ja-En	ニュース	20.1	22.8	**23.5**	+0.7	19.2	21.0	**21.9**	+0.9
WMT21 ニュースタスク En-Ja	ニュース	21.1	21.8	**25.0**	+3.2	21.9	23.1	**24.3**	+1.2
WMT21 ニュースタスク Ja-En	ニュース	19.6	21.5	**22.4**	+0.9	18.1	20.7	**21.3**	+0.6
WMT19 頑健性タスク En-Ja	SNS	12.4	12.5	**14.4**	+1.9	15.6	16.8	**17.3**	+0.5
WMT19 頑健性タスク Ja-En	SNS	11.5	12.3	**12.8**	+0.5	16.0	17.2	**17.7**	+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	20.0	20.6	**21.6**	+1.0
WMT20 頑健性タスク Set2 En-Ja	SNS	12.7	13.0	**14.8**	+1.8	16.4	17.4	**17.9**	+0.5
WMT20 頑健性タスク Set2 Ja-En	SNS	7.9	8.2	**8.6**	+0.4	12.0	12.6	**14.0**	+1.4
IWSLT21 同時通訳タスク En-Ja	TEDトーク	12.5	13.3	**14.5**	+1.2	12.9	14.3	**14.5**	+0.2

目的: JParaCrawl v1.0, v2.0, v3.0 で学習したモデルを様々なテストセットで評価し、翻訳精度の変化を確認
結果 : JParaCrawl v3.0 を使用することで大半のテストセットで大幅な精度向上

翻訳例

入力文	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.

最新の Web をもとに対訳コーパスを作成することで、近年生まれた用語も正しく翻訳できるように