日英機械翻訳を世界に浸透させるために行った3つのこと

森下 睦
NTT コミュニケーション科学基礎研究所

背景

- 機械翻訳研究業界では**標準的なベンチマーク言語対**が存在
  → フランス語 - 英語、ドイツ語 - 英語、中国語 - 英語等
  → 歴史的な経緯や、データ量等が理由で**日英は含まれない**
  → 多くの研究者がこれらの標準的な言語対で実験を行う
- しかし、日本語話者としては**日本語翻訳に取り組みたい**
  → **入出力が読める**というのは大きな**アドバンテージ**
  → **日本語特有の問題**も多く存在
  → 日本語を世界的な**ベンチマーク言語対に取り入れたい**
- 日本語翻訳の研究がより進められるように
  世界中の多くの人に**日本語翻訳に取り組んでもらえる環境を整備**

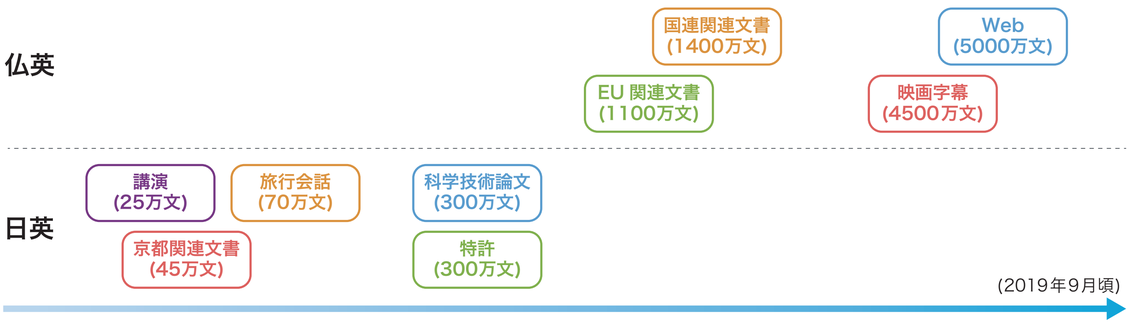
Take Home Message

- 多くの人を巻き込むためには**タスクへの参入障壁を下げる**
  → そのために、データの整備、評価環境の整備が重要
  → その後シェアードタスク化すれば人が集まる
- 多くの人が取り組むことで**タスクの価値が上がる**
  → 自分たちがやっていることの重要性を理解してもらえる
  → 自分たちの論文も通りやすくなる
- 世界中の力を借りて**研究が加速する**
  → 自分たちだけでやれることは限られている
- 皆さんの好きなタスクも環境さえ揃っていれば
  世界中多くの人に興味を持ってもらえるはず

1 実験用データの整備

課題

- 標準的なベンチマーク言語対になるためには、
  誰でも使える**十分なサイズの学習データ**が必要
  → **仏英**などは既に**数千万文規模の対訳コーパス**が存在
  → 一方**日英は大規模な対訳コーパスが無かった**



対応

- Web をクロールし対訳文を集めることで、
  大規模日英対訳コーパス JParaCrawl を構築
  → 現在 2100 万文を超える対訳コーパスに
  → オープンな日英対訳コーパスの中では**世界で最も大きい**
  → データと学習済み翻訳モデルは**一般公開**
- 様々な分野で**機械翻訳精度が大幅に向上**
- 世界中の人が日本語翻訳に取り組みやすい環境に

バージョン	対訳文数	作成時期
v1.0	8,763,995	2019年11月
v2.0	10,120,013	2020年1月
v3.0	21,481,513	2021年12月

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

2 自動評価環境の整備

課題

- 英日翻訳は**自動評価方法が人によってぶれていた**
  → 日本語側の**単語分割が原因**
  → 異なる単語分割 = 異なるスコア = **比較不可能**
  → 評価時に**日本語特有の事情**を考慮する必要があった
- 世界中の人が**誰でも簡単かつ正しく評価できるツール**が必要

対応

- SacreBLEU という**自動評価ツールに日本語対応を実装**
  → 単語分割を内部で行い、公平な比較を可能にするための
  自動評価ツール (初期は英語のみ対応)
- 英日翻訳もこれを使えば**誰でも正しい評価**が可能に
- **1コマンドですぐに評価**できる環境はタスクの普及には重要

A Call for Clarity in Reporting BLEU Scores, Post, WMT18 https://github.com/mjpost/sacrebleu

3 シェアードタスク化

課題

- 日本語翻訳に**取り組める環境は整った**
- しかし日本語翻訳は**日本の組織以外興味がない**状況が続いていた
  → 日本語翻訳に**多くの課題**が残されていることに気づいてほしい
  → 世界中の多くの人に取り組んでもらい**研究を進めてほしい**

対応

- 世界で最も参加者が多い**機械翻訳シェアードタスク WMT** に
  2020 年から**日英・英日翻訳を追加**
- 有名なシェアードタスクの一部となることで、
  多くの人が日本語翻訳に取り組むきっかけに
  → 以降の他の研究論文でも**日英・英日翻訳が使われ始める**

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

Japanese→English

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

2020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

Japanese→English

Rank	Ave.	Ave. z	System
1	73.8	0.141	HW-TSC
2–5	65.1	0.082	IIE-MT
2–6	68.6	0.046	NiuTrans
2–9	67.8	0.033	KwaiNLP
2–6	66.2	0.032	Facebook-AI
5–11	63.5	0.025	XMU
3–10	66.8	0.011	capitalmarvel
5–11	60.9	0.001	Online-B
6–11	61.5	−0.031	MiSS
5–11	66.7	−0.039	Online-W
7–12	59.3	−0.062	WeChat-AI
11–14	59.0	−0.080	Online-A
12–16	55.0	−0.140	Online-G
12–16	64.8	−0.157	movelikeajaguar
13–16	62.2	−0.189	Online-Y
13–16	55.4	−0.193	Illini

2021年

English→Japanese

Range	Ave.	Ave. z	System
1	86.3	0.218	HUMAN-A
2-11	84.1	0.103	NT5
2-9	83.6	0.099	LanguageX
2-9	84.3	0.093	JDExploreAcad.
2-8	84.3	0.087	Online-B
2-9	83.9	0.078	DLUT
2-11	83.2	0.058	Online-Y
3-11	82.9	0.022	Lan-Bridge
6-11	82.9	0.018	Online-A
2-11	83.3	0.004	NAIST-NICT-TIT
11-12	81.9	−0.027	AISP-SJTU
6-12	83.0	−0.029	Online-W
13	79.5	−0.311	Online-G
14	76.9	−0.434	KYB

Japanese→English

Rank	Ave.	Ave. z	System
1	66.7	0.069	DLUT
1	66.1	0.068	NT5
1	66.3	0.059	JDExploreAcademy
1	67.0	0.054	LanguageX
1	68.2	0.049	Online-B
1	66.1	0.046	Online-W
1	68.5	0.016	Lan-Bridge
1	67.1	0.006	Online-G
1	64.8	0.006	Online-A
1	63.8	−0.018	AISP-SJTU
1	66.5	−0.021	NAIST-NICT-TIT
1	66.6	−0.035	Online-Y
1	62.5	−0.056	KYB
14	26.2	−1.285	AIST

2022年