Examining Large Pre-Trained Language Models for Machine Translation: What You Don't Know About It

WMT-2023 @EMNLP2023

Motivations

- Large pre-trained language models (PLMs) are very costly: hardware purchase/lease, ML expertise, training/tuning time, data preparation, evaluation
- Are Extra-large PLMs (*xL*-PLMs) superior to smaller-sized PLMs (*S*-PLMs) toward *domain-specific* Machine Translations (MTs)?
- If not always, in what situations?

Strategies

- Compare two off-the-shelf *xL*-PLMs: Meta-AI's wmt21-dense-24-wide-en-X/X-en (WMT21fb) and NLLB (2022) *vs* one much smaller and well-known *S*-PLM Marian Helsinki
- Two domain specific fine-tuning and testing: automotive commercial, and biomedical/clinical from ClinSpEn2022 (different size of data)

Experimental Settings

- Limited-amount *automotive* in-house data: WMT21fb (4.7 billion parameters) developed for multilingual MT *vs* 618 times smaller Marian (7.6 million parameters)[1]
- **Clinical-domain** test: using 250K pairs fine-tuning data from IBECS after careful cleaning, NLLB-200-distilled (1.3B parameters) [2] *vs* 171 times smaller Marian Helsinki

On Commercial Automotive Data

	Marian	WMT21fb
Before fine-tuning	36.91	47.55
After fine-tuning	48.78	59.92
Gain (↑)	32.16%	26.01%

Table: hLEPOR Metric Scores (https://pypi.org/project/hLepor/)

- The xLPLM wins the scores, though Marian's increasing rate is higher. How about *cost-wise*?

On ClinSpEn Clinical/Biomedical Data

- Three **ClinSpEn-MT** tasks:
- 1) Clinical Cases, EN→ES (**CC**): on 202 COVID-19 clinical case reports;
- 2) Clinical Terms (**CT**), EN←ES: 19K+ parallel terms extracted from biomedical literature and electronic health records (EHRs);
- 3) Ontology Concepts (**OC**), EN→ES: 2K+ parallel concepts from biomedical ontology.
- Evaluations displayed below: clinical-Marian *wins* clinical-NLLB in Task-1 (all metrics), Task-2 (METEOR, ROUGE), and Task-3 (METEOR, COMET, ROUGE) on platform metrics.

Logrus-UoM Team in ClinSpEn-2022

- Clinical-Marian (*S*-PLM) as our official system: ranked the **2nd** on Task-1 (via SacreBLEU, BLEU) and Task-3 (via METEOR, ROUGE) respectively.

How *S*-PLM and *xL*-PLM Perform on Clinical Domain using 250K Pairs of Fine-Tuning?

MT	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
			Clinical-Marian		
Task-I:CC	*38.18*	*0.6338*	*0.4237*	*0.3650*	*0.6271*
Task-II:CT	26.87	*0.5885*	0.9791	0.2667	*0.6720*
Task-III:OC	39.10	*0.6262*	*0.9495*	0.3675	*0.7688*
			Clinical-NLLB		
Task-I:CC	37.74	0.6273	0.4081	0.3601	0.6193
Task-II:CT	*28.57*	0.5873	*1.0290*	*0.2844*	0.6710
Task-III:OC	*41.63*	0.6072	0.9180	*0.3932*	0.7477

Table: Evaluation Scores of Clinical-Marian (*S*-PLM) *vs* Clinical-NLLB (*xL*-PLM) on Three MT Tasks using Fine-Tuned Models.

Bibliography

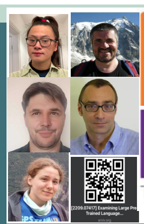
[1] Marcin Junczys-Dowmunt and etc. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*.

[2] NLLB Team. No language left behind: Scaling human-centered machine translation, 2022. URL `https://arxiv.org/abs/2207.04672`.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, Goran Nenadic

{lifeng.han, g.nenadic}@manchester.ac.uk {gleberof, irina.sorokina, serge.gladkoff}@logrusglobal.com

The University of Manchester & Logrus Global LLC | visit https://github.com/HECTA-UoM/ClinicalNMT



LOGRUS GLOBAL

MANCHESTER 1824 The University of Manchester