Investigating Massive Multilingual Pre-Trained Machine Translation Models for Clinical Domain via Transfer Learning

ClinicalNLP2023
@ACL2023

Motivations

- Clinical Knowledge transformation between different languages and regions is significant for supporting global digital healthcare, e.g. creating new quality data and training large language models (LLMs).
- We investigate whether Massively-Multilingual Pre-Trained general domain machine translation (MT) models can be adapted to the clinical domain towards entirely unseen languages via transfer learning.

Strategies

- Take two Massively-Multilingual Pre-Trained general domain MT (MMPMT) models from Meta-AI: 1) NLLB including Spanish as a high resource language, and 2) wmt21-dense-24-wide-en-X/X-en (WMT21fb) without Spanish in its pre-training setting.
- Carry out clinical domain-specific fine-tuning (for NLLB) and transfer-learning (for WMT21fb), then test their performances on English<->Spanish MT.

Models Settings

- Models: NLLB-200-distilled (1.3B parameters) *vs* WMT21fb (4.7 billion parameters).

Fine-tuning Parameters

- batch size = 24
- gradient accumulation steps = 8
- weight decay = 0.01
- learning rate = 2e-5
- number of training epochs = 1
- encoder-decoder layers = 24+24
- Activation function (encoder/decoder) = ReLU

The Parameters for fine-tuning WMT21fb model are the same as for the NLLB-200, except for the batch size which is set as 2, which is because the model is too large and we got an OOM error if the batch size is set above 2.

Fine-tuning using BioMedical Corpus

- 250K pairs of English<->Spanish parallel segments extracted from IBECS after careful cleaning.

Testing On ClinSpEn2022 WMT Data

- Three **ClinSpEn-MT** tasks:
- 1) Clinical Cases, EN→ES (**CC**): on 202 COVID-19 clinical case reports;
- 2) Clinical Terms (**CT**), ES→EN : 19K+ parallel terms extracted from biomedical literature and electronic health records (EHRs);
- 3) Ontology Concepts (**OC**), EN→ES: 2K+ parallel concepts from biomedical ontology.
- Automatic Evaluations + Read the paper for human evaluation examples: https://arxiv.org/abs/2210.06068

Clinical-NLLB (fine-tuning) vs Clinical-WMT21fb (transfer-learning)

MT	SacreBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
			Clinical-WMT21fb		
Task-I:CC	34.30	0.5868	0.3448	0.3266	0.5927
Task-II:CT	24.39	0.5840	0.8584	0.2431	0.6699
Task-III:OC	40.71	0.5686	*0.9908*	0.3859	0.7199
			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 using Five Official Metrics from ClinSpEn2022 Benchmark on Two Models. Clinical-WMT21fb used transfer learning to adapt to a new language Spanish-English, while Clinical-NLLB is normal fine-tuning to the clinical domain.

- Clinical-WMT21fb achieved very close/competitive automatic evaluation scores to Clinical-NLLB.
- Human evaluations further demonstrated that both systems have winning and losing situations in pair-wise comparisons in all three tasks.

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