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

R&D Team of Logrus Global and University of Manchester

For Clinical NLP 2023, July 2023, Toronto

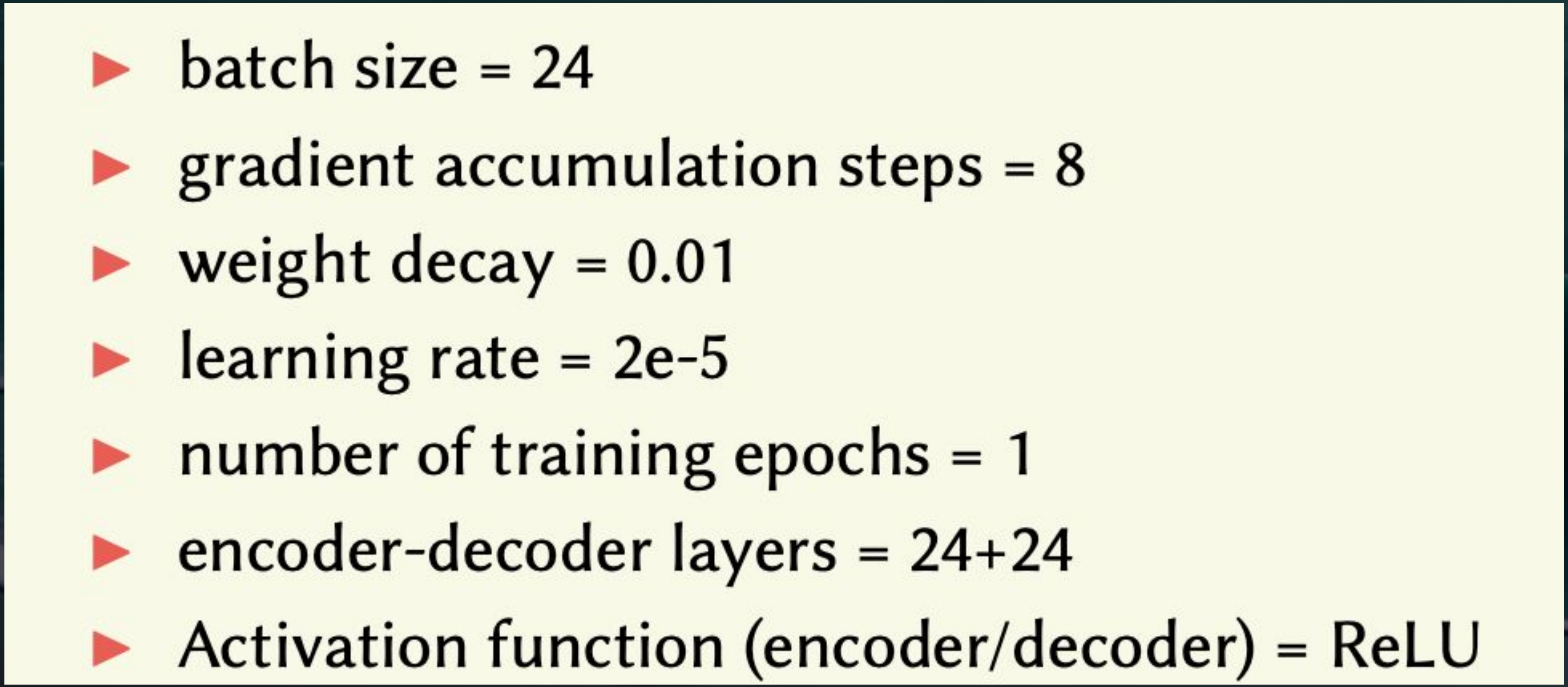
Motivations

- Clinical knowledge transformation is very important across languages for digital healthcare and fine-tuning LLMs in the health domain
- When there is scarce high-quality data available, is it possible to translate the clinical data from one language to another?
- Is transfer learning possible on such MT tasks by fine-tuning LLMs that were not trained using the target/under-study language?
- How to evaluate such transfer-learning results?

Strategies

- We take two multilingual LLMs / LMT models from Meta-AI NLLB and WMT21FB
- NLLB has English-Spanish as a high-resource language pair, while WMT21FB has no Spanish in the training set.
- We take ClinSpEn2022 shared task data that we participated on English-Spanish NMT in clinical domain
- For NLLB we carry out fine-tuning to get clinical-NLLB, while WMT21fb needs a transfer-learning development since the Spanish data / language did not exist, towards the clinical-WMT21fb new model.
- Using <en2ru> and <ru2en> as indicator/pseudo-code for EN->ES and ES->EN towards clinical-WMT21fb
- We evaluate the outputs of these two models with automatic official platform metrics + our human evaluations.

Model Settings



- 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

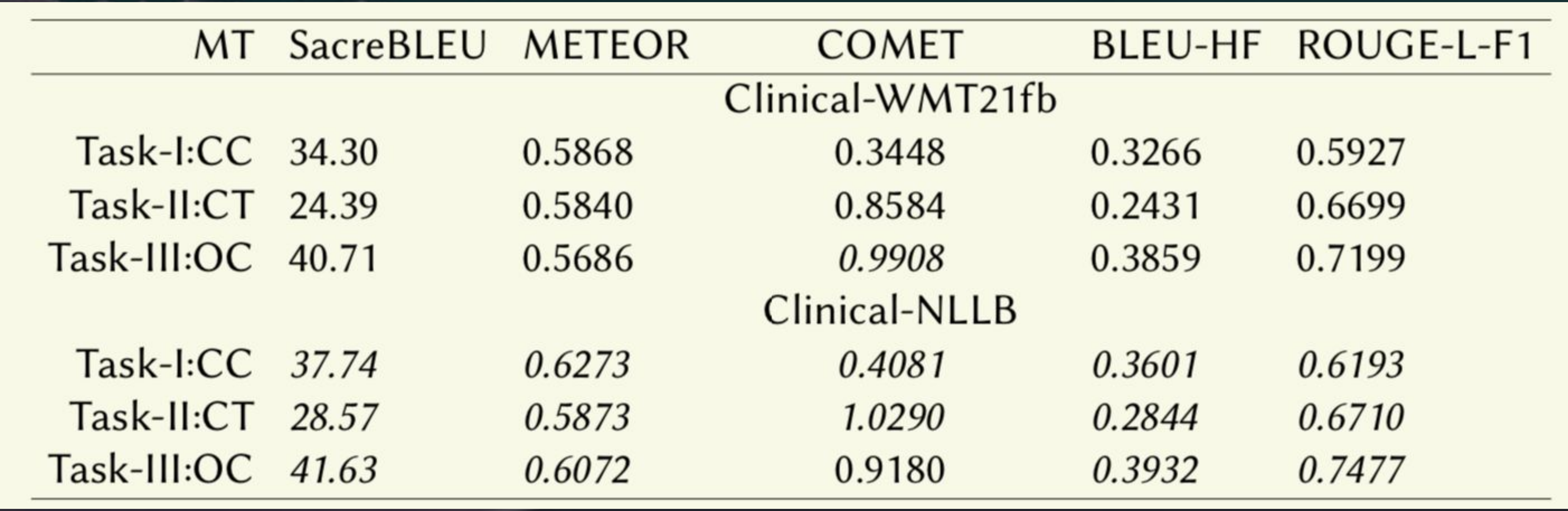
- * NLLB: NLLB-200-distilled (1.3B parameters); the full model is too big
- * WMT21FB: WMT21fb (4.7 billion parameters).

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 and Testing Corpus

- 250 pairs of clean parallel segments in the biomedical domain extracted from IBECS for model fine-tuning
- ClinSpEn2022 shared task data for testing: three tasks on clinical case/report, terms, ontological concepts
- Large amount of testing data to verify the system generalisation

automatic evaluation scores



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*

Findings

- Transfer-learning model Clinical-WMT21FB achieved competitive automatic evaluation scores to Clinical-NLLB-200
- Both fine-tuning and transfer-learning are successful, producing meaningful outputs
- Human evaluation demonstrates both systems have wining outputs in pairwise comparison, and both system needs some post-editing for errors they made
- Clinical knowledge transformation is possible by translating existing clinical text from one language to another, even via LMT models without the language-under-study, as long as sufficient amount of fine-tuning data is available / can be collected.
- Both transfer-learning and fine-tuning system do not reach expert translations in clinical domain, which need further improvement in future work for MT researchers.

Visit github project?

https://github.com/HECTA-UoM/ClinicalNMT

Serge Gladkoff
serge.gladkoff@logrusglobal.com

Lifeng Han
lifeng.han@manchester.ac.uk