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

Lifeng Han[1], **Gleb Erofeev**[2], **Irina Sorokina**[2], **Serge Gladkoff**[2], and **Goran Nenadic**[1]
[1] The University of Manchester, UK
[2] Logrus Global, Translation & Localization
lifeng.han, g.nenadic@manchester.ac.uk
gleberof, irina.sorokina, serge.gladkoff@logrusglobal.com

Abstract

Pre-trained language models (PLMs) often take advantage of the monolingual and multilingual dataset that is freely available online to acquire general or mixed domain knowledge before deployment into specific tasks. Extra-large PLMs (xLPLMs) are proposed very recently to claim supreme performances over smaller-sized PLMs such as in machine translation (MT) tasks. These xLPLMs include Meta-AI's wmt21-dense-24-wide-en-X (2021) and NLLB (2022). *In this work, we examine if xLPLMs are absolutely superior to smaller-sized PLMs in fine-tuning toward domain-specific MTs.* We use two different in-domain data of different sizes: commercial automotive in-house data and **clinical** shared task data from the ClinSpEn2022 challenge at WMT2022. We choose popular Marian Helsinki as smaller sized PLM and two massive-sized Mega-Transformers from Meta-AI as xLPLMs.

Our experimental investigation shows that 1) on smaller sized in-domain commercial automotive data, xLPLM wmt21-dense-24-wide-en-X indeed shows much better evaluation scores using SACREBLEU and hLEPOR metrics than smaller-sized Marian, even though its score increase rate is lower than Marian after fine-tuning; 2) on relatively larger-size well prepared clinical data fine-tuning, the xLPLM NLLB **tends to lose** its advantage over smaller-sized Marian on two sub-tasks (clinical terms and ontology concepts) using *ClinSpEn offered metrics* METEOR, COMET, and ROUGE-L, and totally lost to Marian on Task-1 (clinical cases) on *all official metrics* including SACREBLEU and BLEU; 3) **metrics do not always agree** with each other on the same tasks using the same model outputs; 4) clinic-Marian ranked No.2 on Task-1 (via SACREBLEU/BLEU) and Task-3 (via METEOR and ROUGE) among all submissions.

1 Introduction

Owing to the recent development of neural machine translations (NMTs) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014; Akhbardeh et al., 2021; Han, 2022a), especially the self-attention based Transformer learning structures (Devlin et al., 2019; Vaswani et al., 2017), pre-trained language models (PLMs) have been dominant in natural language understanding (NLU) and natural language processing (NLP) tasks. These applications include Long-Short Term Memory (LSTM) and BERT (Pre-training of Deep Bidirectional Transformers) based models to text mining (Dernoncourt et al., 2017; Wu et al., 2022), question-answering (Dong et al., 2021), reading comprehension (Schlegel, 2021), and summarisation (Perez-Beltrachini and Lapata, 2021), etc., in addition to MT (Han et al., 2021a; Han, 2022b; Han and Gladkoff, 2022).

PLMs often have a large amount of trainable parameters for downstream applications. For instance, in translation task, the popular Marian NMT (Junczys-Dowmunt et al., 2018) pre-trained by Microsoft Translator team [1] on OPUS [2] (Tiedemann, 2012) multilingual corpus has 7.6 million parameters, which can still be fine-tuned on Google's Colab or AWS at virtually no cost. However, very recent work has shown much larger PLMs that have much more parameters than smaller models, e.g. the multi-lingual Transformer model submitted to WMT2021 shared task by Meta-AI research group "wmt21-dense-24-wide-en-x"(WMT21fb) (Tran et al., 2021), which has 4.7 billion parameters, i.e. 618 times bigger than Marian, and does not fit into regular GPUs. In this year, Meta-AI published another model NLLB (NLLB Team et al., 2022) that has 54.5 billion parameters and covers 200 languages in the full model. From now on, we name

[1] https://translator.microsoft.com
[2] https://opus.nlpl.eu

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 908 - 919

both "wmt21-dense-24-wide-en-x" and NLLB as Meta-AI's *Mega-Transformer* models. Meta-AI's Mega-Transformer (WMT21fb) has claimed the best performing system on 10 out of 14 language pairs in WMT2021 shared task including winning bilingual-trained models.

In this work, we raise the question whether extra-large PLMs (xLPLMs) such as Meta-AI's Mega-Transformers have absolute superiority in NMT tasks on domain-specific fine-tuning. We prepare experimental investigation on two different data set to answer this question. One is our specific automotive domain in-house commercial data and the other is clinical domain data from ClinSpEn2022 challenge task we attended which is affiliated with WMT2022 [3].

We set up the following *hypothesis* and *research questions*. Our ***hypothesis*** is: xLPLMs do not absolutely demonstrate superiority over smaller sized PLMs in NMT fine-tuning and it shall depends on specific tasks deployed including domain topic, size of available in-domain data, and performance-cost trad-off.

From this hypothesis we derive two *research questions (**RQs**)*: 1) Do xLPLMs always demonstrate better performances in NMT over smaller sized PLMs for domain fine-tuning? 2) if not, in what situations?

To the best of our knowledge, this is the first published work that has been carried out in the field on fine-tuning Meta-AI's extra-large multilingual PLM Maga-Transformers, and in translating specialised automotive and clinical data.

The rest of the paper is organised as below: Section 2 introduces more details on related work to ours including PLMs and fine-tuning in automotive and clinical domains, Section 3 describes our initial model settings including deployed baseline models, Section 4 presents our experimental evaluation carried out on our in-house commercial automotive domain data, Section 5 describes our system submission to ClinSpEn Biomedical-MT challenge task at WMT2022 on clinical data, and Section 6 gives our conclusion and future work plan.

2 Related Work

Fine-tuning PLMs has been in practice towards different domain applications in recent years. For instance, Wang et al. (2021) carried out experimental investigation on fine-tuning PLMs for conversational recommendation system, Chakraborty et al. (2020); Gu et al. (2021); Lee et al. (2019); Alsentzer et al. (2019) built biomedical and clinical domain pre-trained models using BERT structure and PubMed data on scientific publications, and then Wu et al. (2022); Han et al. (2022a) developed new machine learning structures using PLM Transformer and BERT as encoders in concatenation with statistical graph-based conditional random fields (CRFs) as decoders for clinical text mining.

However, aforementioned work did not deploy extra-large PLMs in a scale as Meta-AI's multilingual Mega-Transformers. For example, the PLMs (Transformer-CRFs) deployed by Wu et al. (2022) as baseline have around 42 million of trainable parameters, which set is already relatively large, even though it is still far from Mega-Transformers' 4.7 billion and 54.5 billion parameters.

Regarding PLM applications in automotive domain, the only recent work we found is from Romell and Curman (2022), who tested the DistilBERT and XLM-RoBERTa PLMs for text classification task using Swedish truck manufacturer data, instead of MT.

There are also researchers working on the overview of model comparability, bench-marking, and fine-tuning methodologies regarding larger scale PLMs, e.g., from Aßenmacher (2021); Ruder (2021).

Overall, none of the work mentioned before has investigated into extra-large Mega-Transformer level PLMs (xLPLMs) for NMT in automotive and clinical domains, especially their comparisons to smaller sized PLMs.

3 Initial Model Settings

To investigate into PLMs with fine-tuning for specialised domain NMT from different scales, we firstly deploy two of such models in different sizes from a multilingual setting. The first one is the popular Marian NMT model developed in C++ since 2018 using deep RNN and Transformer (Junczys-Dowmunt et al., 2018). It is mostly maintained by the Microsoft Translator team and features with efficiency, fast training, and state-of-the-art NMT architectures [4]. This PLM has a smaller sized 7.6 million trainable parameters.

[3] The 7th Conference on MT https://www.statmt.org/wmt22/

[4] available at https://marian-nmt.github.io

The second one we use for fine-tuning is one of the extra-large PLMs (*xLPLMs*) Meta-AI's Mega-Transformer "wmt21.dense-24-wide.En-X" (Tran et al., 2021) developed for WMT2021 shared task on multilingual MT, which was submitted to 14 language pairs and claimed the best on 10 of them [5]. It has 4.7 billion trainable parameters, which is super large in comparison to Marian model. In the later section (5), we will explain another Mega-Transformer Model NLLB developed in this year by Meta-AI and deploy it for our ClinSpEn2022 shared task submission on clinical domain.

4 Model Fine-Tuning and Comparison on Commercial Automotive Data

4.1 In-house Corpus and Hardware

At the development stage, we use our in-house prepared domain-specific commercial corpus from automotive field. We split our data set into 90% vs 10% for fine-tuning and testing respectively and make sure that the test data is not seen during the fine-tuning / development stage [6]. We use a larger GPU from NVIDIA A100 with 80GB VRAM for our experiments because of the much higher computational powers the Mega-Transformer model requires.

4.2 Our Evaluation Setup

BLEU (Papineni et al., 2002) has always been criticised by researchers on its reliability. This includes very recent work by Freitag et al. (2021), which demonstrates that BLEU has closer correlation to lower quality crowd sourced human evaluation then to expert based human evaluation, and by Han et al. (2021a), which investigation on Chinese-English NMT shows that BLEU score fails to reflect the real quality differences between NMT systems especially on translating multi-word expressions (MWEs) and terms (Han et al., 2020).

Furthermore, BLEU scores can be very different caused by configurations, such as tokenisation and normalisation strategies applied to the reference text which can lead to 1.8 margin of difference reported by (Post, 2018). In light of these findings, we adopt two alternative evaluation metrics, i.e. SACREBLEU (Post, 2018) and hLEPOR (Han et al., 2013b; Erofeev et al., 2021; Han et al., 2021b) that we will give further details about.

4.2.1 Revisiting SACREBLEU

SACREBLEU is developed by the work from Post (2018) and is maintained online in its Python version [7]. The author discussed the uncertainty regarding reporting BLEU scores by MT researchers. This is involved in many parameter settings when using BLEU metric including number of references, length penalty computation on multi-references, maximum n-gram, and smoothing applied to 0-count n-grams. Because of such variations, when MT researchers report the BLEU scores from their system, "the BLEU" score actually cannot be reproduced in many cases due to lack of detailed technical description of encoder, etc. .

To address these issues, SACREBLEU added some constrains while using BLEU metric. These include the applying of its own metric-internal pre-processing for detokenised system outputs, the avoiding of user handling reference set via automatically downloading from WMT, and the export of a summary on settings used.

4.2.2 Revisiting hLEPOR

hLEPOR is an augmented metric for automatic MT evaluation which was firstly proposed in WMT2013 Metrics shared task (Han et al., 2013a,b) and was reported as one of the best performing metrics at both system level (Macháček and Bojar, 2013) and segment level (Graham et al., 2015) [8]. It is calculated via a weighted harmonic mean of several main factors including sentence length penalty, position difference penalty, precision, and recall. Furthermore, there are more weighting parameters among all the sub-factors. Let's see the brief formulas below:

$$hLEPOR = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR)$$

where *LP* is the sentence length penalty factor and is calculated as:

[5] package "wmt21.dense-24-wide.En-X" available at https://github.com/facebookresearch/fairseq/tree/main/examples/wmt21

[6] Because this is a commercial corpus, we do not give much details on it but this does not affect the experimental findings we achieved

[7] available at https://github.com/mjpost/sacrebleu

[8] The python version is available at https://pypi.org/project/hLepor/ and the original Perl code at https://github.com/poethan/LEPOR

$$\text{LP} = \begin{cases} e^{1-\frac{Length_{ref}}{Length_{hyp}}} \; if \; Length_{hyp} < Length_{ref} \\ 1 \; if \; Length_{hyp} = Length_{ref} \\ e^{1-\frac{Length_{hyp}}{Length_{ref}}} \; if \; Length_{hyp} > Length_{ref} \end{cases}$$

Then, n-gram based position difference penalty (NPD) is used to measure the word position and order difference among matched words between system output and reference translation ($MatchN_{hyp}$ and $MatchN_{ref}$).

$$NPosPenal = e^{-NPD}$$

$$NPD = \frac{1}{Length_{hyp}} \sum_{i=1}^{Length_{hyp}} |PD_i|$$

$$|PD_i| = |MatchN_{hyp} - MatchN_{ref}|$$

Finally, the weighted harmonic mean of precision and recall is calculated using this formula.

$$HPR = \frac{(\alpha + \beta) PrecisionxRecall}{\alpha Precision + \beta Recall}$$

$$Precision = \frac{Aligned_{num}}{Length_{hypothesis}}$$

$$Recall = \frac{Aligned_{num}}{Length_{reference}}$$

hLEPOR is an extended version of the original LEPOR metric (Han et al., 2012; Han, 2014). hLEPOR also has a latest customised version named cushLEPOR which uses automatic hyper-parameter optimisation framework Optuna (Akiba et al., 2019) to achieve better and easier feature weights fine-tuning towards specific language pairs and domains in practice. It was reported as one of the best performing metrics in WMT2021 (Erofeev et al., 2021; Han et al., 2021b) on the officially-ranked language pairs English-German and Chinese-English on News domain, and English-Russian on TED talk data (Freitag et al., 2021) where human expert level evaluations were available. hLEPOR is also gaining popularity in other NLP task evaluations, e.g. language generation (NLG) (Novikova et al., 2017; Gehrmann et al., 2021; Marzouk, 2021), language understanding (NLU) (Ruder et al., 2021), text summarization (ATS) (Bhandari et al., 2020), and searching (Liu et al., 2021).

4.3 Evaluation Results

The evaluation scores using SACREBLEU and hLEPOR are shown in Table 1 and 2 respectively. From Table 1, we can see that the fine-tuning has successfully improved each single n-gram precision score in SACREBLEU for both Marian and Mega-Transformer models, leading to an overall 150.14% and 75.81% score increasing. Similarly, Table 2 shows that our in-domain fine-tuning improved hLEPOR scores on Marian and Mega-Transformer models via 32.16% and 26.01%.

Like BLEU, SACREBLEU is precision based metric. The very large margin evaluation score increases in SACREBLEU (150.14% and 75.81%) indicates that according to reference translation, our fine-tuned models produce more fluent output than the baseline in this domain specific test set. Unlike SACREBLEU, hLEPOR is an augmented metric with comprehensive factors, including recall and positional difference penalty, in addition to precision. The large margins of hLEPOR score increase, i.e. 32.16% and 26.01% tell that the fine-tuned models can also have more adequate translation towards this domain, in addition to maintaining higher fluency.

In summary, the fine-tuning of these two PLMs has demonstrated evaluation score improvement with large margins in commercial domain data. xLPLM Mega-Transformer has much higher SACREBLEU evaluation score than Marian before fine-tuning, 39.12 vs 19.64, which indicates its larger amount of knowledge acquired. However, after fine-tuning, the SACREBLEU scores of them are much closer, 50.33 vs 45.20. This means that fine-tuning of smaller sized PLM for this commercial data is far more effective than the xLPLM Mega-Transformer from computation and time cost point of view, as well as the cost of computational power itself, since supercomputer time is much more expensive.

This partially verifies our assumption that xLPLMs do not always win smaller sized PLMs in practical applications when computational cost is in place and when time is constrained.

To further investigate our research questions, we carry out another experimental evaluation on clinical domain data via attending the ClinSpEn2022 shared task challenge which will be detailed in the next section.

	Marian					
	uni-gram	bi-gram	tri-gram	4-gram	BP	Overall
Before fine-tuning	19.64	10.96	4.56	2.00	1.0	7.38
After fine-tuning	45.20	24.54	14.44	8.69	0.96	18.46 (↑150.14%)
	Mega-Transformer (wmt21fb)					
	uni-gram	bi-gram	tri-gram	4-gram	BP	Overall
Before fine-tuning	39.12	18.81	9.78	5.23	1.0	13.93
After fine-tuning	50.33	30.14	19.47	12.85	0.99	24.49 (↑75.81%)

Table 1: SACREBLEU score comparisons on the MT test set: before vs after fine-tuning

	Marian	Mega-Transformer
Before fine-tuning	36.91	47.55
After fine-tuning	48.78	59.92
Rate(↑)	32.16%	26.01%

Table 2: hLEPOR score comparisons on the MT test set: before vs after fine-tuning

5 Submission to ClinSpEn at WMT22

In this section, we introduce our system submissions to Biomedical-MT task in WMT2022. In this task, we attended the affiliated clinical domain machine translation on Spanish-English language pair (ClinSpEn) task [9], which is hosted in CodaLab (Pavao et al., 2022) [10].

The aim of this task is to promote the development of MT models on medical domain via three sub-tasks: 1) Clinical Cases (CC): on 202 COVID-19 clinical case reports; 2) Clinical Terms (CT): using more than 19K parallel terms extracted from biomedical literature and electric health records (EHRs); 3) Ontology Concepts (OC): using more than 2K parallel concepts from biomedical ontology. The translation direction on these three sub-tasks are EN→ES, EN←ES, and EN→ES respectively.

5.1 Corpus Used

In addition to the official corpora prepared by the ClinSpEn organisers, we used some external corpora for our model fine-tuning. This is because that neural-network based machine learning models are data dependent while the officially offered parallel sample sentences are very limited. We found useful biomedical Spanish-English corpora described in (Névéol et al., 2018) from WMT[11], and MeSpEn corpora from (Villegas et al., 2018)[12], which include Spanish Bibliographical Index in Health Sciences (IBECS), Scientific Electronic Library Online (SciELO), and U.S. National Library of Medicine (PubMed and MedlinePlus). However, due to the time restriction for this shared task, we only managed to get 250,000 aligned pairs from IBECS after careful preparation, which is a bibliographical data collecting scientific articles from different fields of health sciences, maintained by the Spanish National Health Sciences Library.

5.2 Adaptations on xLPLM: NLLB

Two systems we submitted to ClinSpEn2022 are clinic-Marian and clinic-NLLB (NLLB Team et al., 2022). We reported our clinic-WMT21fb model outputs in a followup work (Han et al., 2022b) (also due to the time restriction). Some training parameters and training logs for clinic-Marian are listed below:

- batch size = 64
- gradient accumulation steps = 1
- weight decay = 0.01
- learning rate = 2e-5
- number of training epochs = 1
- number of examples = 225,000

NLLB (No Language Left Behind) is another extra-large PLM model built by Meta-AI freshly in

[9] https://temu.bsc.es/clinspen/
[10] https://codalab.lisn.upsaclay.fr/competitions/6696
[11] https://github.com/biomedical-translation-corpora
[12] https://zenodo.org/record/3562536

this year [13], which was targeting low-resource languages via knowledge transfer from high-resource ones, and Spanish is among the high-resource languages covered by NLLB [14]. NLLB-200 has a total of 54.5 billion parameters in its full model as the authors mentioned. In this shared task, we applied the distilled version of NLLB, i.e. the "NLLB-200-distilled-1.3B" which still has 1.3 billion trainable parameters [15]. As Meta-AI's "wmt21.dense-24-wide.en-X" model we used in the earlier section, we call NLLB-distilled as one of their *Mega-Transformers*.

Some fine-tuning parameters for NLLB-distilled are listed below:

- 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

The fine-tuned clinic-NLLB model has relatively apparent evaluation score increase using SACREBLEU in comparison to baseline model on both translation directions, as shown in Table 3, for EN→ES and ES→EN in the upper and middle parts of the table with increasing rate 11.74% and 9.70% respectively. This demonstrates that that fine-tuning was successful.

Interestingly, if we fine-tune the model in one direction and carry out the inference translation in the opposite direction, the model performance will have a big drop even though it is the same language pair. This tells that pre-trained LMs lose their generalisation after fine-tuning. For instance, in the bottom of Table 3, we demonstrate that if the model is fine-tuned in English-to-Spanish direction and the inference test is carried out in Spanish-to-English direction, the overall SACREBLEU score has a 14.37% drop in comparison to without fine-tuning. So, we carried out fine-tuning on both translation directions for the system submission to three sub-tasks at ClinSpEn2022.

[13]The project page https://ai.facebook.com/research/no-language-left-behind/

[14]Models available at https://huggingface.co/docs/transformers/model_doc/nllb

[15]https://huggingface.co/facebook/nllb-200-distilled-1.3B

5.3 Official Evaluation Metrics

The official evaluation metrics used by CinSpEn2022 shared task are METEOR (Banerjee and Lavie, 2005), SACREBLEU (Post, 2018), COMET (Rei et al., 2020), BLEU-HF (HuggingFace) (Papineni et al., 2002), and ROUGE-L-F1 (Lin, 2004). Among these, METEOR is a metric using both precision and recall not only on word surface level but also introducing paraphrasing features. COMET was proposed recently by taking advantage of cross-lingual PLMs using knowledge from both source and target languages. ROUGE was originally designed for text summarisation evaluation using n-gram co-occurrences, while ROUGE-L added the Longest Common Sub-sequence (LCS) feature from translation study.

5.4 Evaluation Scores on Three Tasks

We present the MT evaluation scores using five official metrics through CodaLab platform on the three sub-tasks in Table 4, for translating clinical cases, clinical terms, and ontology concepts. The two fine-tuned models are clinic-Marian and clinic-NLLB (one of the Mega-Transformers). From this shared task evaluation outcomes, the xLPLM clinic-NLLB starts to lose its comparisons to far smaller-sized clinic-Marian in Task-2 (CT) and 3 (OC), especially on METEOR and ROUGE-L scores but also on COMET (OC). What is very noticing is that clinic-Marian has an overall win on Task-1 (CC) via all evaluation metrics.

From the evaluation results on Task 2 and 3, i.e. CT and OC, we can see that the evaluation metrics do not agree with each other always. For instance, clinic-Marian wins METEOR and ROUGE-L on Task 2 but loses on other metrics, while clinic-NLLB wins SACREBLEU and BLEU-HF on Task 3 but loses on other metrics. This phenomenon is very interesting which tells that variation metrics from BLEU including BLEU-HF and SACREBLEU tend to not agree with other metric families including METEOR, COMET, and ROUGE-L. Furthermore, the same metric does not always agree with itself on different tasks, or the two MT models perform differently across tasks. For instance, COMET score says clinic-Marian and clinic-NLLB wins Task 3 (0.9495) and 2 (1.0290) respectively. Due to the time restriction from this shared task and the limited computational resource we have, our second model (clinic-NLLB) was submitted after the official deadline.

	English-to-Spanish (tune+test)					
	uni-gram	bi-gram	tri-gram	4-gram	BP	Overall
Before fine-tuning	65.93	45.51	33.71	25.44	1.0	40.05
After fine-tuning	70.25	50.58	38.78	30.17	0.99	44.75 (↑11.74%)
	Spanish-to-English (tune+test)					
	uni-gram	bi-gram	tri-gram	4-gram	BP	Overall
Before fine-tuning	65.36	42.54	30.58	22.60	1	37.23
After fine-tuning	68.51	46.27	34.07	25.76	1	40.84 (↑9.70%)
	English-to-Spanish (tune) & Spanish-to-English (test)					
	uni-gram	bi-gram	tri-gram	4-gram	BP	Overall
Before fine-tuning (es2en)	65.36	42.54	30.58	22.60	1	37.23
After *Reverse* fine-tuning	58.17	36.48	25.85	18.84	1.0	31.88 (↓14.37%)

Table 3: SACREBLEU score comparisons using NLLB: baseline vs fine-tuned in clinical domain.

This experimental investigation shows that with carefully prepared and larger amount of domain specific data for fine-tuning, the xLPLMs tend to lose its advantage over smaller sized PLMs using several automatic metrics. Thus it further verifies our hypothesis and research questions.

5.5 Comparisons to Other Teams

In the officially valid submissions (before the shared task deadline ended) for three tasks, there are four teams for Task-1 and Task-3 including Avellana Translation, DtranX, Optum and ours [16]. In addition to these four teams, Task-2 has another team Huawei, making it in-total five teams. Optum and Huawei have both multiple submissions/runs while other teams submitted one run. Our submission clinic-Marian ranked number 2 in both Task-1 and Task-3 via SACREBLEU/BLEU and METEOR/ROUGE respectively, as in Table 5 underlined. There are four runs from Optum team for both Task-1/3 and single submission by other teams. Table 5 includes the best submission from Optum. There is a little difference in the last digit of the evaluation scores between our own record (Table 4) and the official record (Table 5), which is because that we rounded the last digit scores while the official ones did not. This result shows that metrics tend to not agree with each others in many cases. For instance, on Task-3, our clinic-Marian has very similar score to DtranX on METEOR (0.6261 vs 0.6275) only from the third digit which is a metric using paraphrase and semantic similarity features; however, the score difference on BLEU is so large (39.10 vs 58.24) via SACREBLEU which rises the issue again on the credibility of BLEU metric. There are not many teams submitting their results into this clinical domain machine translation task in comparison to the traditional news domain MT task, which indicates that it is still a relatively new domain and calls for more attentions from MT researchers in the future.

6 Discussion and Future Work

In this work, we carried out experimental investigations on if extra-large pre-trained language models (PLMs) always demonstrate superiority over much smaller-sized PLMs using two domain specific data. The first experimental results using Marian vs "wmt21.dense-24-wide.En-X" shows that even though xLPLM still perform better evaluation scores in comparison to much smaller sized Marian, their score difference is much smaller after fine-tuning and the xLPLM costs more than smaller PLM from performance-cost trade-off point of view in practical applications, e.g. for language service providers (LSPs). The second experimental results using clinical data show that with carefully prepared certain amount of fine-tuning data (250k sentence pairs), the xLPLM NLLB even loses with its evaluation score in comparison to smaller PLM Marian in Task 1 "clinical cases" over all automatic metrics used, and in Task 2 "clinical terms" and 3 "ontology concepts" on partial of the automatic evaluation metrics officially used by ClinSpEn2022. Finally, our system submission clinic-Marian ranked the second place using SACREBLEU/BLEU for

[16] https://statmt.org/wmt22/biomedical_results.pdf

	clinic-Marian				
MT	SACREBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Task-I: clinical cases	*38.18*	*0.6338*	*0.4237*	*0.3650*	*0.6271*
Task-II: clinical terms	26.87	*0.5885*	0.9791	0.2667	*0.6720*
Task-III:clinical concepts	39.10	*0.6262*	*0.9495*	0.3675	*0.7688*
	clinic-NLLB (Mega-Transformers)				
MT	SACREBLEU	METEOR	COMET	BLEU-HF	ROUGE-L-F1
Task-I: clinical cases	37.74	0.6273	0.4081	0.3601	0.6193
Task-II: clinical terms	*28.57*	0.5873	*1.0290*	*0.2844*	0.6710
Task-III: ontology concepts	*41.63*	0.6072	0.9180	*0.3932*	0.7477

Table 4: Evaluation Scores using Official CodaLab Platform from ClinSpEn2022 Benchmark on Fine-tuned Models. *italic* scores indicate winner on the specific task using the specific metric (last digit rounded).

	Task-1: Translating Clinical Cases				
Teams	SACREBLEU	METEOR	COMET	BLEU	ROUGE
DtranX	*41.06*	0.6633	*0.4610*	*0.3926*	*0.6490*
Logrus-UoM (ours)	38.17	0.6337	0.4237	0.3650	0.6270
Optum(run4)	38.12	0.6447	0.4425	0.3642	0.6285
Avellana Translation	36.64	*0.6637*	0.3920	0.3519	0.6333
	Task-3: Translating Ontology Concepts				
Teams	SACREBLEU	METEOR	COMET	BLEU	ROUGE
DtranX	*58.24*	*0.6275*	*1.2496*	0.5724	*0.7839*
Optum(run4)	44.97	0.5880	1.1197	0.4396	0.7479
Logrus-UoM (ours)	39.10	0.6261	0.9494	0.3674	0.7688
Avellana Translation	31.72	0.5707	0.3841	0.3042	0.7621

Table 5: Comparisons on Task 1 and 3 across teams (ranked via SACREBLEU chosen by the organisers).

Task-1, and using METEOR/ROUGE for Task-3 among all teams who submitted on-time before the shared task deadline.

We looked into the translation outputs from clinic-NLLB for error analysis, and it shows that some of the translation errors come from very literal translation, and others come from gender related mistakes. In conclusion, *our two stage experimental investigations verify our hypothesis and RQs from different aspects*.

We also doubt if the official automatic metrics used for ClinSpEn challenge can correctly distinguish the NMT systems because mostly they do not really measure the translation output quality but the similarity to the gold standard single reference. Therefore, domain specific automatic evaluation metrics or metrics better measuring semantic similarities might be needed.

In the future work, we plan to carry out more experimental investigations from qualitative aspects looking into translation errors using human experts and classifying them into possible categories with examples and statistics, especially from clinical domain. This will allow us to validate automatic metrics with professional human judgements for this domain.

We will continue to fine-tune our models towards different domains and languages and use more of the available corpus for current clinical domain challenge task. We also plan to try different state-of-the-art pre-trained language models for evaluation.

Acknowledgements

The authors thank the ClinSpEn2022 shared task organisers for preparing the data set and evaluation platforms, thank Darryl Estrada for in communication with us during the competition. The authors

thank the project support from HIPS (R126035 task A05) and from JigSaw (R124782 task A07) at The University of Manchester. We thank the open research projects Marian Helsinki NMT and Meta-AI's wmt21.dense-24-wide.En-X (2021) and NLLB (2022) we used.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, OndÅ™ej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina EspaÃ±a-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthias Aßenmacher. 2021. *Comparability, Evaluation and Benchmarking of large pre-trained language models*. Ph.D. thesis, Ludwig Maximilian University of Munich.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhe Dong, Guangrun Wang, Hang Xu, Jiefeng Peng, Xiaozhe Ren, and Xiaodan Liang. 2021. EfficientBERT: Progressively searching multilayer perceptron via warm-up knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1424–1437, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gleb Erofeev, Irina Sorokina, Lifeng Han, and Serge Gladkoff. 2021. cushLEPOR uses LABSE distilled knowledge to improve correlation with human translations. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 421–439, Virtual. Association for Machine Translation in the Americas.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *arXiv e-prints*, page arXiv:2104.14478.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Six Conference on Machine Translation*. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D.

Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. *arXiv e-prints*, page arXiv:2102.01672.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013a. A description of tunable machine translation evaluation systems in WMT13 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 414–421, Sofia, Bulgaria. Association for Computational Linguistics.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau, MSc. Thesis.

Lifeng Han. 2022a. *An investigation into multi-word expressions in machine translation*. Ph.D. thesis, Dublin City University.

Lifeng Han. 2022b. An overview on machine translation evaluation. *arXiv preprint arXiv:2202.11027*.

Lifeng Han, Valerio Antonini, Ghada Alfattni, Alfredo Madrid, Warren Del-Pinto, Judith Andrew, William G. Dixon, Meghna Jani, Ana Maria Aldana, Robyn Hamilton, Karim Webb, and Goran Nenadic. 2022a. A transformer-based machine learning framework using conditional random fields as decoder for clinical text mining. In *HealTAC 2022: the 5th Healthcare Text Analytics Conference. Poster 19.*

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022b. Using massive multilingual pre-trained language models towards real zero-shot neural machine translation in clinical domain.

Lifeng Han and Serge Gladkoff. 2022. Meta-evaluation of translation evaluation methods: a systematic up-to-date overview. In *Tutorial at LREC2022*, Marseille, France.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. 2021a. Chinese character decomposition for neural MT with multi-word expressions. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 336–344, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021b. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1014–1023, Online. Association for Computational Linguistics.

Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013b. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.

2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-evaluation of Conversational Search Evaluation Metrics. *arXiv e-prints*, page arXiv:2104.13453.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Shaimaa Marzouk. 2021. An in-depth analysis of the individual impact of controlled language rules on machine translation output: a mixed-methods approach. *Machine Translation*.

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Alv Romell and Jacob Curman. 2022. *Multilingual Large Scale Text Classification for Automotive Troubleshooting Management*. MSc Thesis, Lund University.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. *arXiv e-prints*, page arXiv:2104.07412.

Viktor Schlegel. 2021. *Emerging evaluation paradigms in natural language understanding: a case study in machine reading comprehension*. Ph.D. thesis, University of Manchester.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai's wmt21 news translation task submission. In *Proc. of WMT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: multilingual biomedical text processing*.

Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph. *CoRR*, abs/2110.07477.

Yuping Wu, Lifeng Han, Valerio Antonini, and Goran Nenadic. 2022. On cross-domain pre-trained language models for clinical text mining: How do they perform on data-constrained fine-tuning? In *Forthcoming*.

Appendix

More training logs from clinic-Marian:

- global step = 3516
- training loss = 1.2236216656855212
- train runtime = 1945.9989
- train samples per second = 115.622
- trian steps per second = 1.807
- total flos = 2947034863632384.0

Parameters reported by SACREBLEU:

- lowercase = Ture
- tokenize = 13a