Using Massive Multilingual Pre-Trained Language Models Towards Real Zero-Shot Neural Machine Translation in Clinical Domain

Lifeng Han¹, Gleb Erofeev², Irina Sorokina², Serge Gladkoff², and Goran Nenadic¹

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

Massively multilingual pre-trained language models (MMPLMs) are developed in recent years demonstrating super powers and the pre-knowledge they acquire for downstream tasks. In this work, we investigate whether MMPLMs can be applied to zero-shot machine translation (MT) towards entirely new language pairs and new domains. We carry out experimental investigation using Meta-AI’s MMPLMs “wmt21-dense-24-wide-en-X and X-en (WMT21fb)” which were pre-trained on 7 language pairs and 14 translation directions including English to Czech, German, Hausa, Icelandic, Japanese, Russian, and Chinese, and opposite direction. We fine-tune these MMPLMs towards English-Spanish language pair which did not exist at all in their original pre-trained corpora both implicitly and explicitly. We prepare carefully aligned clinical domain data for this fine-tuning, which is different from their original mixed domain knowledge as well. Our experimental result shows that the fine-tuning is very successful using just 250k well-aligned in-domain EN-ES pairs / sentences for three sub-task translation testings: clinical cases, clinical terms and ontology concepts. It achieves very close evaluation scores to another MMPLM NLLB from Meta-AI, which included Spanish as a high-resource setting in the pre-training. To the best of our knowledge, this is the first work on using MMPLMs towards real zero-shot NMT successfully for totally unseen languages during pre-training, and also the first in clinical domain for such study.

1 Introduction

Multilingual neural machine translation (MNMT) has its root from the beginning of NMT era (Dong et al., 2015; Firat et al., 2016) but only made its first milestone when Google’s end-to-end MNMT arrived (Johnson et al., 2017) where the artificial token was introduced for the first time for translation task at the beginning of the input source sentence to indicate the specified target language, e.g. “2en” as translating into English. This model used a shared word-piece vocabulary and enabled multilingual NMT through a single encoder-decoder model training. Google’s MNMT also demonstrated the possibility of “zero-shot” translation as long as the languages to be translated from or to have been seen during the training stage, even though not explicitly. However, as the authors mentioned, Google’s MNMT only allows to translate between languages that have been individually as “source and target languages during some point, not for entirely new ones” in their many-to-many model, which was tested using the WMT14 and WMT15 data (Johnson et al., 2017). This set an obstacles to translate freshly new languages that do not exist in their pre-training stage. Then using the later developed NMT structure Transformer and BERT (Devlin et al., 2019; Vaswani et al., 2017), Facebook AI extended the coverage of multilingual translation into 50, 100, and 200+ languages via mBERT-50 (Tang et al., 2020), M2M-100 (Fan et al., 2021), and NLLB (NLLB Team et al., 2022) models. However, these models never address the issue on translating entirely new languages that do not exist in their pre-training stage, which sets an obstacles for MT applications in serving an even broader community.

In this work, we move one step forward towards real zero-shot NMT via fine-tuning an entirely new language pair that does not exist in the deployed multilingual pre-trained language models (MPLMs). The MPLMs we used are from Facebook AI (Meta-AI)’s submission to WMT21 news translation task, i.e. “wmt21-dense-24-wide-en-X” and “wmt21-dense-24-wide-X-en” which were pre-trained for 7 languages Hausa (ha), Icelandic (is), Japanese (ja), Czech (cs), Russian (ru), Chinese (zh), German (de) to English (en), and backwards (Tran et al., 2021). We use a well prepared 250k
pairs of English-Spanish (en-es) clinical domain corpus and demonstrate that not only it is possible to achieve real zero-shot translation on this explicit new language pair, i.e. the Spanish language is totally unseen among the languages in the MPLM, but also the domain knowledge transfer from general and mixed domain to clinical domain is very successful. In comparison to the massively MPLM (MMPLM) NLLB which covers Spanish as a high resource language at its pre-training stage, our zero-shot model achieves very close evaluation scores in most sub-tasks (clinical cases and clinical terms translation) and even wins NLLB in ontology concept translation task on the metric COMET (Rei et al., 2020) using ClinSpEn2022 testing data at WMT22.

3 Model Settings

To investigate into our RQ, we take Meta-AI’s MNMT submission to WMT21 shared task on news translation, i.e. the MMPLM “wmt21-dense-24-wide-en-X” and “wmt21-dense-24-wide-X-en” as our test-base, and we name them as WMT21fb models (Tran et al., 2021). They are conditional generation models from the same structure of massive M2M-100 (Fan et al., 2021) having a total number of 4.7 billion parameters which demand high computational cost for fine-tuning. WMT21fb models were trained on mixed domain data using “all available resources” they had, for instances, from historical WMT challenges, large-scale data mining, and their in-domain back-translation. Then these models were fine-tuned in news domain for 7 languages including Hausa, Icelandic, Japanese, Czech, Russian, Chinese, German from and to English.

The challenge language we choose is Spanish, which does not appear in the training stage of WMT21fb models. The fine-tuning corpus we use is from MeSpEn (Villegas et al., 2018) clinical domain data, of which we managed to extract 250k pairs of English-Spanish sentences. They are from IBECS-descriptions, IBECS-titles, MedlinePlus-health_topics-titles, MedlinePlus-health_topics-descriptions, Pubmed-descriptions, Scielo-descriptions, and Scielo-titles, and we carried out data cleaning.

To implement the fine-tuning, we use the <2en> token for translating from Spanish to English, and <2ru> (originally to Russian) pseudo token 2 using <2es> token will result into errors since Spanish was actually not used in the WMT21fb PLMs.

2 Related Work

Regarding the early usage of special token in NMT, Sennrich et al. (2016) designed the token T from Latin Tu and V from Latin Vos for familiar and polite indicators attached to the source sentences towards English-to-German NMT. Yamagishi et al. (2016) designed tokens <all-active>, <all-passive>, <reference> and <predict> to control of voice of Japanese-to-English NMT; either they are active, passive, reference aware or prediction guided. Subsequently, Google’s MNMT system designed target language indicator, e.g. <2en> and <2jp> controlling the translation towards English and Japanese respectively (Johnson et al., 2017). Google’s MNMT also designed mixed target language translation control, e.g. (1-alpha)<2ko>+alpha<2jp> tells a mixed language translation into Korean and Japanese with weighting mechanism. We take one step further to use an existing language controller token from a MPLM as a pseudo code to fine-tune an external language translation model, which is entirely not seen during pre-training stage.

Regarding zero-shot applications for downstream NLP tasks other than MT, Muller et al. (2021) applied transfer learning from MPLMs towards unseen languages of different typologies on dependency parsing (DEP), named entity recognition (NER), and part-of-speech (POS) tagging. Ahuja et al. (2022) carried out zero-shot transfer learning for natural language inference (NLI) tasks such as question answering.

In this paper, we ask this research question (RQ): Can Massive Multilingual Pre-Trained Language Models Create A Knowledge Space Transferring To Entirely New Language Pairs and New Domains for Machine Translation In A Zero Shot Fashion?
In comparison, we deploy another MMPLM from Meta-AI, i.e. the “No-Language-Left-Behind (NLLB)” which was trained on 204 languages including Spanish as one of their high-resource ones (NLLB Team et al., 2022). NLLB full model is a massive size Transformer having 55 billion parameters and we use its distilled version NLLB-200-distilled 3, which still has 1.3 billion parameters. Fine-tuning is carried out on NLLB using the same 250K ES-EN corpus.

4 Model Evaluations

4.1 Testing Corpus from Clinical Domain

We use the official testing corpus from ClinSpEn2022 shared task affiliated to Biomedical-MT at WMT22. ClinSpEn2022 aims at developing clinical domain machine translation on Spanish-English language pair4, which is hosted in Codalab (Pavao et al., 2022)5.

There are 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 electronic 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.

4.2 Evaluation Metrics

The official evaluation metrics used by ClinSpEn2022 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 Subsequence (LCS) feature from translation study.

The reporting of BLEU metric scores has certain uncertainty, which is caused by some 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. 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.

4.3 Evaluation Scores

We present the MT evaluation scores using five official metrics from ClinSpEn2022 shared task on the three sub-tasks in Table 1, for translating clinical cases, clinical terms, and clinical concepts. The two fine-tuned models are clinic-NLLB which is not zero-shot and clinic-WMT21fb which is a zero-shot NMT model.

On Task 1 and 2, Clinical-WMT21fb has very comparable evaluation scores to clinical-NLLB,
Table 1: Evaluation Scores using Five Official Metrics from ClinSpEn2022 Benchmark on Two Models.

| Task-I: Clinical Cases (CC) EN→ES | MT fine-tuning | zero-shot? | SACREBLEU | METEOR | COMET | BLEU-HF | ROUGE-L-F1 |
|----------------------------------|----------------|------------|------------|--------|-------|---------|------------|
| Clinical-NLLB                    | No             | 37.74      | 0.6273     | 0.4081 | 0.3601| 0.6193  |            |
| Clinical-WMT21fb                 | Yes            | 34.30      | 0.5868     | 0.3448 | 0.3266| 0.5927  |            |

| Task-II: Clinical Terms (CT) EN←ES |
|-----------------------------------|
| MT fine-tuning | zero-shot? | SACREBLEU | METEOR | COMET | BLEU-HF | ROUGE-L-F1 |
| Clinical-NLLB | No         | 28.57      | 0.5873  | 1.0290| 0.2844 | 0.6710    |
| Clinical-WMT21fb | Yes       | 24.39      | 0.5840  | 0.8584| 0.2431 | 0.6699    |

| Task-III: Ontology Concept (OC) EN→ES |
|--------------------------------------|
| MT fine-tuning | zero-shot? | SACREBLEU | METEOR | COMET | BLEU-HF | ROUGE-L-F1 |
| Clinical-NLLB | No         | 41.63      | 0.6072  | 0.9180| 0.3932 | 0.7477    |
| Clinical-WMT21fb | Yes       | 40.71      | 0.5686  | 0.9908| 0.3859| 0.7199    |

even though it only used 250k pairs es-en sentences for fine tuning without seeing any en-es or Spanish language at all during pre-training. In contrast, clinical-NLLB used a large amount of Spanish data for its pre-training phase. On Task 3, the evaluation scores of these two models are even closer on BLEU and SACREBLEU, especially the zero-shot model clinical-WMT21fb winning COMET metric over clinical-NLLB (0.9908 vs 0.9180).

This experimental result shows that with carefully prepared certain amount of fine-tuning data, e.g. 250k pair of sentences, the MMPLMs are capable create a semantic knowledge space transferring to an entirely new / external language pair for NMT task in a new domain, e.g. clinical domain. This answers our RQ set up in the beginning of this investigation.

4.4 Human Evaluation

We look into three sub-task translation output files from the zero-shot model clinical-WMT21fb. It shows that for the EN←ES translation task, i.e. the sub-task 2 clinical term translation, the output file is totally file with only English tokens. On the other two sub-tasks, i.e. the clinical cases and ontology concept translation, which have the translation direction EN→ES, there are some Russian tokens in the output, not only Spanish tokens. However, the Russian tokens in the Spanish sentences are not nonsense, instead proper translations of entities and words. The entire test set of these two sub-task is very large around 300K sentences/lines, and there are only 12K lines of them (4%) have Russian tokens. So we have fine-tuned the model in EN-RU direction on EN-SP data, and it translates well into Spanish! But if there isn’t suitable Spanish token in the training data, it takes Russian token.

We also looked into the translation outputs from clinic-NLLB model for error analysis, and it shows that some of the translation errors come from very literal translation, and others come from gender related mistakes. This suggests that the massively pre-trained MLM is still not there to capture the difference of linguistic features among pre-trained languages.

5 Discussion

5.1 On Automatic Metrics

We had more thoughts on the automatic evaluation settings and outputs, especially on the COMET metric in comparison to others.

Firstly, the closeness of most automatic metric scores do not necessarily mean that the translation outputs are very good. Most metrics only measure the linguistic proximity of outputs to the “gold standard of reference”.

Secondly, COMET is a reference-less metric taking advantage of cross-lingual PLMs using knowledge from both source and target languages. This has pros and cons: a) it might be able to capture the semantic relatedness without seeing the same language tokens, even in the same sequence/sentence; b) also due to this, it is not able to distinguish foreign language tokens in the translation output, which normally shall receive penalty in evaluation scores. This also inspires another research topic, i.e. shall we really punish the foreign or mixed-
language tokens in the translation output in all evaluation conditions, or it shall be depends on situation of the output applications? This has an echo to Google’s zero-shot MNMT model (Johnson et al., 2017) when the mixed language tokens are used for translation model, e.g. \((1-\alpha)\langle2KO\rangle + \alpha\langle2JP\rangle\) resulting in mixed tokens of Korean and Japanese in the output translation but they are semantically correct tokens.

In a situation when users want only the Spanish translation output, 4% of Russian tokens in the Spanish translation should surely receive a penalty in quality evaluation setting. The COMET metric will fail this mission, and professional human evaluation is always much needed for trust worthiness. However, in a situation to measure the models’ cross-lingual capability on semantic preservation for direct output, or as input into another ML models, is it better to generate NULL or meaningless tokens or random translations in the target language, or to choose semantically correct foreign tokens when the model does not know how to predict the exact correct target tokens? This inspires us to think again on the evaluation setting on different tasks.

5.2 On PLM Capability for Zero-Shot

In this work, we used Meta-AI’s WMT21 multilingual pre-trained language models as our test-base for the knowledge transfer into an external language fine-tuning and translation. This zero-shot ability is much dependent on the MPLMs we used, such as WMT21fb (Tran et al., 2021) as a huge size model, a conditional generation from Meta-AI’s massive M2M-100 model (Fan et al., 2021). If we try to fine-tune a bilingual model on an external language which the PLM did not see, it will not be that good because for smaller sized models such fine-tuning would be too much of a change, and the model will lose generalisation which leads to problems. For huge multilingual PLM models, the 250K of fine-tuning data is small set of number, and that’s why the model does not lose generalisation and captures new data well without losing linguistic knowledge of other languages that it was trained on.

6 Conclusions and Future Work

We investigated if real zero-shot NMT is possible using massive multilingual pre-trained LMs (MM-PLMs) to translate external languages that are unseen at all in the training phase. We used Meta-AI’s mixed domain multilingual PLMs (WMT21fb) as our test-base, 250K well prepared EN-ES clinical data as fine-tuning corpus, and \(<2ru>\) as pseudo code for new language (out-of-en) fine-tuning. We tested the fine-tuned zero-shot model on ClinSpEn2022 clinical domain shared task data, and the results show that this fine-tuning is successful, which achieves very comparable scores to Meta-AI’s MMPLM NLLB model, which had Spanish in the training phase as high-resource setting. We think this demonstrates that the Hyper-Transformer model from WMT21fb does build a language independent “semantic space” that allows to understand a different language and correctly construct a totally different language model when fine-tuned on the language which is absent and different from the languages it was trained upon. We will try more external languages from different typologies in future work.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

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

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.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1723–1732, Beijing, China. 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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40, San Diego, California. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. CoRR, abs/2008.00401.

Chau Tran, Shrutii 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, Aitor Inxtaurondo, 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.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Appendices

Parameters

Some fine-tuning parameters for NLLB-200-distilled (NLLB Team et al., 2022) 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

Some fine-tuning parameters for WMT21fb Models (Tran et al., 2021) are listed below:
• batch size = 2 (note\(^6\))
• gradient accumulation steps = 8
• weight decay = 0.01
• learning rate = 2e-5
• Activation function (encoder/decoder) = ReLU
• number of training epochs = 1
• encoder-decoder layers = 24+24

More details on M2M-100 for Conditional Generation structure (Fan et al., 2021) we used can be find in Figure 2.

\(^6\)the model is too large that we get OOM error if batch size is set above 2
Figure 2: M2M-100 Model Structure For Conditional Generation Encoder Sample.